A FORMANT–BASED LINEAR PREDICTION SPEECH SYNTHESIS/ANALYSIS SYSTEM

BY

YEAN–JEN SHUE

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1995

To my parents,

my wife, Mei–Ing, and my sons, Ian and Will

ACKNOWLEDGMENTS

I wish to express my gratitude to my supervisor and committee chairman, Dr. D. G. Childers, for his invaluable guidance, encouragement, and patience throughout the past four years of my graduate study.

I would like to express my sincere appreciation to Dr. F. J. Taylor, Dr. J. C. Principe, Dr. H. B. Rothman, and Dr. J. M. M. Anderson for serving on my supervisory committee and providing invaluable suggestions for my research. I am grateful to my colleagues at the Mind–Machine Interaction Research Center for their help and friendship. Special thanks go to Dr. J. M. White for his suggestions and patience. I would also like to thank Mrs. S. Ashwell for helping me correct the grammar in my dissertation.

Thanks also go to the Chung Shan Institute of Science and Technology for granting me the scholarship to pursue the Ph.D. degree.

Last but not least, I am deeply indebted to my parents, my wife Mei–Ing, and my sons Ian and Will for their love, support, and understanding. My greatest gratitude is to them.

TABLE OF CONTENTS

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment
of the Requirements for the Degree of Doctor of Philosophy

A FORMANT–BASED LINEAR PREDICTION SPEECH SYNTHESIS/ANALYSIS SYSTEM

By

Yean–Jen Shue

May, 1995

Chairman: Dr. D. G. Childers
Major Department: Electrical Engineering

The aim of this research was to develop a speech synthesis/analysis system as the framework for generating high–fidelity synthetic speech and for psychoacoustic studies. A formant–based linear prediction (LP) synthesizer, along with a robust speech analysis procedure, was developed to achieve this aim. The major feature of this system is its ability to adapt the formant and linear prediction schemes to represent the voiced and unvoiced sounds, respectively. The advantages of employing two kinds of schemes in one synthesis system are 1) the formant scheme is physically meaningful for simulating the human speech production system, and 2) the LP scheme is able to reproduce the spectrum of all speech sounds.

The formant–based LP synthesizer uses two types of sources, voiced and unvoiced, to form the excitation part of the synthesizer. These sources are either nonparametric waveforms or parametric models of waveforms. The vocal tract is characterized by a twelfth order linear prediction filter. For voiced sounds, the coefficients of the vocal tract filter are determined by the first six formants. The counterparts for unvoiced sounds are obtained by means of a twelfth order LP analysis. This synthesizer can resynthesize speech almost perfectly when the estimated glottal waveform from a glottal inverse filtering process is used as the excitation source. When the modeled waveform is used as the excitation source, the synthesized speech is natural and intelligible.

The other feature of this research is that the interaction between the synthesis and analysis is closely defined. A two–phase, LP–based procedure that analyzes a segment of the speech signal was developed to estimate the time–varying synthesis parameters such as the voiced/unvoiced classification, fundamental frequency, signal power, formants (for voiced sounds), LP coefficients (for unvoiced sounds), and the estimated glottal waveforms to the formant–based LP synthesizer.

Based on the synthesis and analysis procedures, as well as a knowledge of the relationships between vocal quality and glottal features, a voice conversion procedure that reproduces the vocal tract component, but varies the glottal features, was developed to convert a segment of the speech signal of modal voice type to five other voice types (vocal fry, breathy, falsetto, whisper, and harsh). The conversion procedure provides a systematic method for examining the relationships between vocal quality and glottal features, and can be used to build a data base for various voice types, which can be used in training a speech recognition system.

In addition to the glottal source parameters, the vocal tract parameters can be manipulated by our synthesis/analysis system as well. Since the features of the glottal source and the vocal tract are both involved in speech studies such as gender conversion, speaker identification, and speech recognition, this synthesis/analysis system can serve as a tool for future applications.

CHAPTER 1
INTRODUCTION

1.1 Speech Synthesis and its Applications

Speech synthesis is the process of creating a synthetic replica of a speech signal. The first major effort of speech synthesis was made by Wolfgang Von Kempelen (1791) who invented a mechanical talking machine, which could speak whole phrases in French and Italian. During the nineteenth century, a number of studies were made to investigate the generation process of vowel sounds by using mechanical devices, and these results strongly supported the evolution of speech research (reviewed by Linggard, 1985). After 1930, along with the advent of electronic instrumentation such as the oscilloscope, people began to understand speech acoustics and use this knowledge along with electronic circuits to build speech synthesizers. Dudley's (1936) ten–channel vocoder was the first truly electronic speech synthesizer. Owing to its flexibility, the electronic speech synthesizer outperformed the conventional mechanical synthesizers of the time. In 1960, the availability of the digital computer made it possible to adopt software programs instead of hardware to do the synthesis. This evolution not only allowed speech scientists to implement and evaluate a proposed speech synthesizer without really having to build it by hardware, but also encouraged the applications of speech synthesis.

Speech communication is one kind of application for speech synthesis. High synthetic speech quality (intelligible and/or natural) and low bit–rate (less bandwidth)

transmission are the main concerns for this application (Flanagan et al., 1980). In order to satisfy the goal, the design of this type of synthesis scheme usually models the speech signal itself rather than the mechanism of human speech production, and hence it is relatively easier to design. Successful examples such as the glottal excited linear prediction (GELP) synthesizer and the sinusoidal coders have been well studied (Childers and Hu, 1994; Rose and Barnwell, 1990; Trancoso et al., 1990).

Another application of speech synthesis is for psychoacoustic studies (Rabiner and Schafer, 1978; Eggen, 1992; Eskenazi et al., 1990), which can be thought of as one kind of analysis–synthesis technique — by varying the synthesis parameters (of a specific synthesizer) in a controlled fashion and assessing the resulting synthetic speech through a subjective response (listening test) or objective measure (spectral distance), speech scientists are able to learn certain phenomena of the human speech production process (Klatt, 1980; Lalwani, 1991; Childers and Lee, 1991). Instead of modeling the speech signal, this kind of synthesis design is used to simulate the human speech production system either physically or acoustically.

In fact, such synthesis schemes for psychoacoustic studies have already been proposed. Linggard (1985) said that the articulatory synthesizer, which is designed to physically model the movements and positions of articulators (the larynx, jaw, tongue, velum, and lips) during speech production, has the potential to explore certain articulatory and acoustic features of the human speech production system. Other studies (Klatt, 1980; Holmes, 1983; Pinto et al., 1989; Lalwani, 1991) have introduced the advantages of applying the formant synthesizer to psychoacoustic applications. Recent research showed that the modern linear prediction techniques have been utilized to extract certain selected acoustic cues for the physiological study of the vocal folds (Hu, 1993; Childers and Hu, 1994; Gavidia–Ceballos and Hansen, 1994). However, because of the immaturity of the analysis–synthesis process (articulatory synthesizer), the improper modeling of the unvoiced sounds (formant synthesizer), and a lack of resemblances between the human

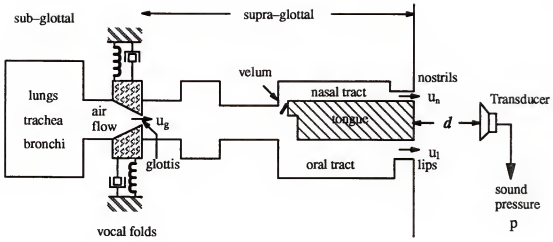
speech production process and the synthesis model (LPC synthesizer), the use of these known synthesizers for psychoacoustic studies is restricted. This restriction motivates us to develop a synthesis scheme that is capable of not only producing high–fidelity synthesized speech, but also being a useful tool for exploring certain essential features of the speech production process.

1.2 Literature Review

1.2.1 Speech Production Mechanism

Since we are interested in studying the possibility of building a synthesis system that is able to simulate the human speech production mechanism for psychoacoustic studies, it will be necessary to review this mechanism first.

Figure 1–1(a) illustrates a simplified schematic diagram of the vocal apparatus (Rabiner and Schafer, 1978; Kaplan, 1971). The lungs, trachea, and bronchi constitute the sub-glottal system from which the air flow is generated (Miller, 1959). The glottis is known as the opening between the vocal folds. The amount of the air flow that passes through the glottis during a unit time is defined as the glottal volume-velocity u_g. This flow excites the supra-glottal system, which is made up of two coupled tubes. The oral tube begins at the glottis and ends at the lips, and the nasal tube begins at the velum and ends at the nostrils. The position of the velum determines whether the nasal tube and oral tube are coupled or not. Both tubes are resonant cavities. For each cavity, their natural resonant frequencies are determined by the shape of the cavity, the viscous friction between the air and the walls of the cavity, and the heat conduction coefficient of the walls of the cavity. Mathematically, the cavity can be characterized by the resonances. The glottal volume-velocity is filtered by the supra-glottal system and generates an output volume-velocity at the orifices (lips and



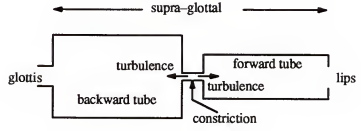
u_g : volume–velocity at the glottis

u_l : volume–velocity at the lips

u_n : volume–velocity at the nostrils

p : sound pressure measured by a transducer at a distance ***d***

(a)



(b)

Figure 1–1. Human speech production mechanism.
(a) Schematic diagram of vocal apparatus;
(b) Schematic diagram for the production of unvoiced sounds.

nostrils). The quantities, u_L and u_n, are defined as the volume-velocity at the lips and nostrils, respectively. The sound pressure, p, is included in the speech production process in order to simulate the human hearing system. This is because sound pressure is the measurable quantity that stimulates the aural membrane.

In American English, there are about 42 phonemes that constitute the basic elements of speech (Rabiner and Schafer, 1978). According to the characteristics of the excitation source, these phonemes can be classified into three broad categories: voiced, unvoiced, and mixed sounds. Various mechanisms are used for generating various categories of sounds.

In generating voiced sounds, air flow from the sub-glottal system is modulated by the vibrating vocal folds, and results in a quasi-periodic, pulse-like excitation (glottal volume–velocity, u_g). This waveform, which interacts with the supra–glottal system, is closely correlated with various voice types, which will be discussed later (Childers and Lee, 1991; Childers and Ahn, 1994). For non–nasalized voiced sounds, since the velum is in closed position, there is no air flow in the nasal tube, and the supra-glottal system can be modeled as an all-pole system. In producing nasal phonations, the velum is opened, and the oral tube is constricted at the lips. The coupled nasal and oral tubes will form anti–resonances, and thus the supra–glottal system is a pole–zero system (Klatt, 1980). The characteristics of the supra-glottal system for voiced sounds are considered to be one of the key factors for gender identification and conversion (Murry and Singh, 1980; Kuwabara and Ohgushi, 1984; Childers and Wu, 1989; Wu, 1990).

In generating unvoiced sounds, the glottis is open with the glottal volume-velocity constant. A narrow constriction is formed in the oral tube. According to the aerodynamic theory, this constriction will generate a turbulence flow. For various types of unvoiced phonations (fricatives, stops, and affricates), the acoustic features and duration of this turbulence flow are still research topics (Stevens et al., 1992; Stevens, 1993a; Stevens, 1993b). Because of the constriction, the oral tube becomes two tubes that are excited by the turbulence flow as shown in Figure 1–1(b). The forward tube starts at the constriction

and ends at the lips and the backward tube starts at the constriction and ends at the glottis. The coupling effect between these two tubes makes it reasonable to consider the vocal tract as a pole-zero system.

For mixed sounds, the supra-glottal system is similar in manner to that for unvoiced sounds, except that the vocal folds vibrate. The superposition process of the excitations (glottal volume–velocity and turbulent flow) is used to explain the production mechanism for mixed sounds. The power ratio between the glottal volume–velocity and the turbulent flow is approximately 10–15 dB (Stevens, 1992).

1.2.2 Speech Synthesizer

The speech production mechanism can be simulated either by an articulatory or an acoustic model, and thereby the speech synthesis schemes have been designed accordingly (Fant, 1960; Flanagan et al., 1975).

1.2.2.1 Articulatory synthesizer

The articulatory synthesizer that originates from Kempelen's work (1791) attempts to mechanically model the positions and movements of the articulators. The modern designs have adopted electrical concepts to simulate the excitation sources and the vocal tract.

In modeling the glottal source for voiced sounds, the cross–sectional area of the glottis is the essential variable. Since it is changed relatively faster than the other variables, more complicated algorithms such as the two–mass model have been proposed to describe the dynamic phenomenon of the vibrations of the vocal folds (Sondhi, 1975; Flanagan and Ishizaka, 1978; Titze, 1982; Chan, 1989). The parameters of these complicated source models are useful to interpret the physiological and pathological phenomena of the vocal

folds. For unvoiced sounds, the turbulent flow is generated by applying constant glottal volume–velocity to the constriction inside the oral tube. The cross–sectional area and the shape of this constriction determine the acoustic features of the turbulent flow.

In simulating the vocal tract, the cross–sectional area function, A(x), is often specified at 20 to 40 separate points along the oral tube. Research has been done to derive the area function for various phonemes either by traditional X–ray photography or by acoustic optimization procedures (Fant, 1960; Gopinath and Sondhi, 1970; Levinson and Schmidt, 1983; Hsieh, 1994). The former method is reliable but tedious. The latter method inverse filters the speech waveform, requires considerable computation and does not guarantee a satisfactory result. The movement of the articulators for connected speech causes A(x) to be not only position–dependent but also time–varying. In a discrete realization, the cross–sectional area function, A(x), is updated every 20–50 msec. The interpolations of the area function in between two specified points and in between two specified instants are crucial for synthesizing high–quality speech (Schroeter et al., 1987; Schroeter et al., 1988; Linggard, 1985).

Recent studies have adopted an analogue transmission line, as shown in Figure 1–2, to simulate the basic blocks (excitation source, vocal tract, and radiation filter) of the articulatory synthesizer (Flanagan and Ishizaka, 1978). The sub-glottal pressure, P_s, is assumed to be constant. The glottal impedance, Z_g, made up by the series connection of the glottal resistance, R_g, and inductance, L_g, are determined by the vibration model of vocal folds. The oral tract is modeled as a sequence of RLC circuits in cascade. The values of these R_n, L_n, C_n elements are determined by their corresponding cross–sectional area, A_n. The radiation filter (transforming the volume-velocity to sound pressure) at the lips is approximated by a parallel combination of a loading resistance R_L and inductance L_L. As long as the cross–sectional area function is appropriately specified, this type of articulatory synthesizer promises high quality synthesized speech as well as considerable

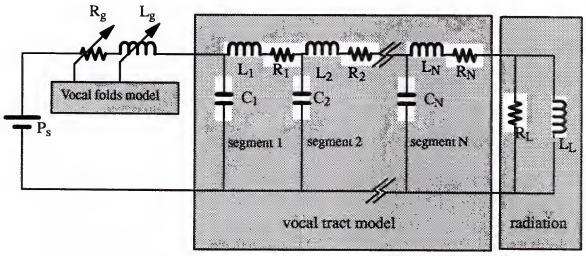


Figure 1–2. The articulatory synthesizer, a circuit model.

insight into the dynamics of the speech production process (Linggard, 1985; Gupta and Schroeter, 1993).

1.2.2.2 Formant synthesizer

A simplified approximation of the speech production mechanism in the acoustic domain was proposed in the late 1950s and called the source–tract or source–filter model (Fant, 1960). In this model, the speech production system is split into two parts: 1) the excitation source and 2) the resonant tract. These two parts are noninteractive and linearly connected. The formant synthesizer is one example that applies the source–tract model as the synthesis scheme. Formants (frequency, bandwidth, and intensity), the natural resonances of the supra–glottal system, are adopted by speech scientists as a more meaningful representation in modeling the acoustic features of the vocal tract.

Since it is an acoustic model, the history of the formant synthesizer is closely related to the evolution of electronic technology. In the late 1930s, "terminal–analog" synthesizers were made by using analogue electrical networks. These analogue networks are serial or parallel combinations of second order resonators. A series of impulse–like waveforms, or white noise is applied to the resonators in order to generate the vowel–like or fricative sounds respectively.

In the 1960s, the discrete domain realizations began to make the study of formant synthesizers flourish (Flanagan et al., 1962; Rabiner, 1968). Instead of analogue resonators, second order digital resonators are used in the discrete implementation. By comparing the resonance pattern of the uniform acoustic tube, with a five pole analogue and a five pole digital resonator, Fant (1956) indicated that a compensating network is needed by the analog filter for the absence of higher formants (Fant, 1956), and this compensating network is not required for the digital implementations (Gold and Rabiner, 1968). This is because the periodic nature of the digital resonator effectively adds in higher

poles to its transfer function, thus the extra correction process is not necessary. This definitely is another advantage of adopting a digital formant synthesizer.

Since the resonators can be arranged in either a cascade or parallel manner, there is an argument as to which is the better configuration. Flanagan (1957) concluded that the serial type is the better model for non–nasal voiced sounds, while the parallel structure is superior for nasal and unvoiced sounds. The reason is that the vocal tract is considered as an all–pole filter for non–nasal voiced sounds and as a pole–zero system for other phonations, and it is quite simple to use the cascade structure to simulate the all–pole system and the parallel one to implement the pole–zero system. In 1980, Klatt proposed a formant synthesizer that combined the cascade and the parallel structures together, as shown in Figure 1–3. This synthesis system is proposed to obtain benefits from both structures. Anti–resonators have been added to the cascade branch in order to enhance the ability of using the cascade configuration to model the nasal or unvoiced sounds. By properly specifying the synthesis variables and using the correct configuration, this synthesizer is capable of synthesizing highly intelligible speech.

As is well known, the synthesis process can never stand alone; it needs controls, correct synthesis variables, etc. For the formant synthesizer, the formants are the most important variables that affect the synthesis quality. Therefore, there are numerous procedures that have been used to derive accurate values for the formants. Most of these procedures use the acoustic speech signal as the source for determining the formants (Olive, 1971; Markel and Gray, 1974; McCandless, 1974; Klatt and Kaltt, 1990; Alku, 1992; Childers and Lee, 1991). Unfortunately, almost all of these procedures have only addressed the extraction processes for vowel–like sounds. Nevertheless, some of the literature has listed the formants for certain phonemes as shown in Table 1–1 (Klatt, 1980). It is difficult to measure the value of the first formant by inspecting the frequency response of an unvoiced fricative / s /, as shown in Figure 1–4 (b). However, the formants for a sustained vowel, such as / i / shown in Figure 1–4 (a), can be measured in several ways.

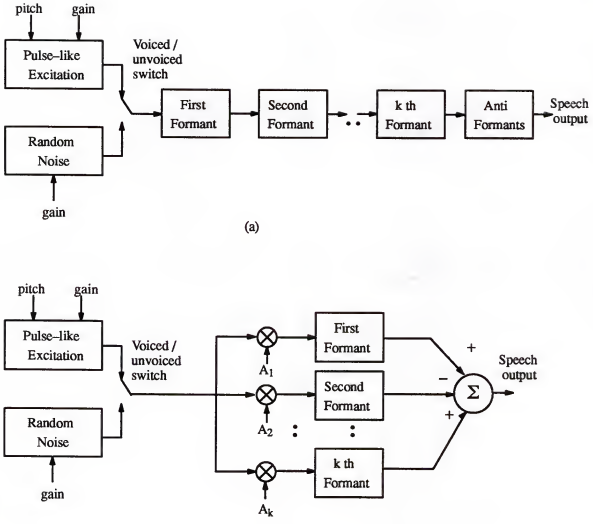
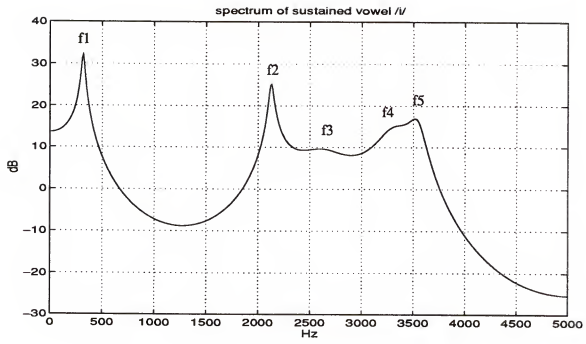
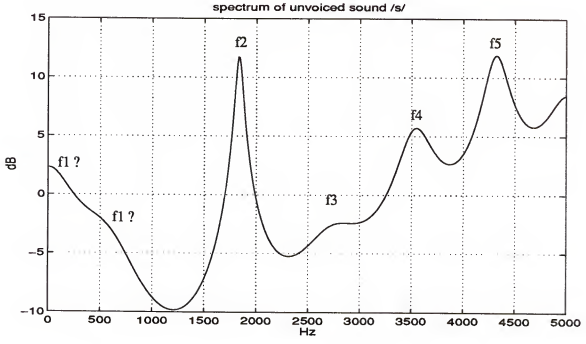


Figure 1–3. Schematic diagrams of a formant synthesizer.
(a) Cascade configuration;
(b) Parallel configuration.



(a)



(b)

Figure 1–4. Frequency response of one period of speech signals. Frequency response is obtained by LPC method. (a) Sustained vowel / i /; (b) Unvoiced sound / s /.

Therefore, even though research has adopted the formant synthesizer for the synthesis–by–rule application by creating a table of the formants for each phoneme (Klatt and Klatt, 1990), the formant synthesizer has not been widely accepted as a research tool.

Table 1–1. Formant frequency, bandwidth for the selected phonemes (Klatt, 1980)

phoneme	f1	f2	f3	b1	b2	b3
s	320	1390	2530	200	80	200
i	310	2020	2960	45	200	400

1.2.2.3 LP synthesizer

The linear predictive (LP) synthesizer is a mathematical realization of the linear source–tract model. The linear prediction process has been used to estimate the poles of a system or of a signal. The basic idea behind this process is that a signal sample, s(n), can be estimated by a linear combination of its past signal samples s(n–k)

$$\hat{s}(n) = \sum_{k=1}^{p} a_k \, s(n-k) \qquad (1-1)$$

where $\hat{s}(n)$ is the estimated signal at instant n, and p is the order of the linear predictor. The linear predictive coefficients, a_k, are determined by minimizing the total error, E, which is the sum of the squared differences, e(n), for a sequence of N samples

$$E = \sum_{n=1}^{N} e^2(n) \qquad (1-2)$$

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{P} a_k\, s(n-k) \qquad (1-3)$$

By transforming Eq.(1–3) to the Z–domain and rearranging it, we obtain

$$S(z) = \frac{1}{1 - a_k \sum_{k=1}^{P} Z^{-k}}\ E(z) = V(z)\, E(z) \qquad (1-4)$$

$$V(z) = \frac{1}{1 - a_k \sum_{k=1}^{P} Z^{-k}} \qquad (1-5)$$

where V(z) is an all-pole transfer function.

By applying the above idea to speech signals, the block diagram of an LP synthesizer is formed and shown in Figure 1–5 (Markel and Gray, 1974). Two types of excitation sources are switched at the input of the all-pole system because of the acoustic differences between them. In generating voiced sounds, the excitation source is pulse–like and governed by the pitch period and gain parameter. For unvoiced sounds, the excitation source, E(z), is generated by a random noise, which is controlled by the gain parameter and spectral parameter. The transfer function, V(z), is characterized by the LP coefficients, a_ks, and implemented by a time-varying digital filter. Since it is more like a mathematical model, the parameters controlling the process for the LP synthesizer is completely automatic and does not require formant tracking and a spectral fitting procedures. Because of its simplicity, the all–pole system is more popular than the pole–zero system in modeling speech signals. Furthermore, this model has proven to be quite satisfactory in practice.

Traditional LP synthesizers will produce pop and click sounds because of the poor modeling process used for the excitation source. Recent research has shown that various LP synthesizers, such as the coded excited linear predictive (CELP) synthesizer, can

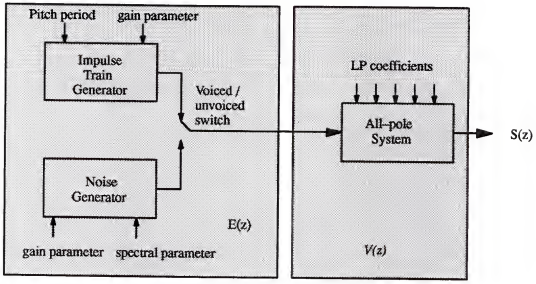


Figure 1–5. Block diagram of a typical LP synthesizer.

reproduce high quality speech (Childers and Hu 1994; Milenkovic, 1993; Trancoso et. al., 1990).

There are many advantages (such as high–fidelity and ease of control) for applying the LP scheme for synthesis and analysis. However, since the LP technique can not acoustically decompose the speech signals into an excitation source and a vocal tract, the LP scheme is not suitable for the physiological study of the human speech production system. The reason for the unsuccessful separation process is that the LP minimization process has assumed the error, e(n), to be random noise, while research has shown that the glottal excitation source is a periodic pulse–like waveform for voiced sounds (Fujisaki and Ljungqvist, 1986; Olive, 1992).

Olive (1992) has proposed a mixed spectral representation (formant and LP) to make use of the benefits of both the formant and the LP synthesis schemes. Basically, his strategy is to use the high order LP scheme for synthesizing unvoiced sounds and the extended formant scheme for voiced phonations. By carefully considering the discontinuity problem that might arise at the boundary of voiced and unvoiced sounds, Olive (1992) claims that his synthesizer can be used for analysis–synthesis applications and produce high–quality synthesized speech.

1.2.3 Vocal Quality and Glottal Excitation Source

Vocal quality is a term that has many meanings (Laver, 1980; Laver and Hanson, 1981), but is usually referred to as the total auditory impression the listener experiences upon hearing a speech token (Childers, and Lee, 1991; Gobl and Chasaide, 1992). Laryngeal vocal quality has been defined as the aspects of vocal quality related to the vibratory pattern of the vocal folds (Van den Berg, 1968). Throughout this paper we will use the term "vocal quality" to represent the laryngeal vocal quality. Because of this

definition, the speech tokens that are being inspected by listeners (physicians or speech experts) usually are confined to voiced sounds.

Laver and Hanson (1981) classified vocal quality into six major types: modal, vocal fry, breathy, falsetto, whisper, and harsh. Sometimes we refer to these six types of vocal quality as six voice types. Modal is defined as the normal mode of phonation. Vocal fry is perceptually a low pitch, rough sounding phonation. A breathy voice contains a slightly audible friction. Falsetto is perceived as a flute-like tone that is sometimes breathy. Whisper is perceived as strongly audible friction. Harsh is an unpleasant, rough, rasping sound. To a certain extent, the voice types such as breathiness and harshness are thought of as the symptoms of a malfunction of the vocal folds. In other words, vocal quality can be applied to the physiological or pathological studies of the vocal folds.

A subjective evaluation of the human voice is the general way of accessing vocal quality. However, because subjective tests are dependent upon the training and experience of the assessor, it may not be adequate to rely on the use of the subjective method to reach a conclusion about vocal quality (Gavidia–Ceballos and Hansen, 1994).

Numerous studies have examined the possibility of applying objective measures as an auxiliary tool to analyze the relationships between voice types and certain acoustic features of the glottal source (Matsumoto et al., 1973; Singh and Murry, 1978; Kasuya et al., 1986; Childers and Lee, 1991; Childers and Ahn, 1994). By analyzing speech and electroglottographic (EGG) signals of four voice types (modal, breathy, vocal fry, and falsetto), Childers and Lee (1991) reported that four factors (glottal pulse width, glottal pulse skewness, the abruptness of glottal closure, and the turbulence noise) are significant for characterizing the glottal source. Preliminary perceptual tests were also conducted in their experiment to evaluate the auditory effects of the glottal source parameter upon synthesized speech that was generated by a formant synthesis process. Childers and Ahn (1994) used the glottal inverse filtering procedure and the statistical analysis tool (analysis of variance, ANOVA) to quantitatively examine the relationships between glottal source

(LF model) parameters and three voice types (modal, breathy, and vocal fry). In the research mentioned above, the voice types of the speech tokens were pre–evaluated by qualified experts.

Since it is difficult to derive the acoustic features of the glottal source by direct analysis of the speech signal for various voice types, such as whisper and harsh, research has turned to the analysis–synthesis technique to study the causes for various voice types (Kuwabara, 1991; Childers and Lee, 1991; Lalwani, 1992). The formant synthesizer is the most popular synthesis scheme for this application. By hypothesizing that the glottal source is the main cause of the variation of vocal quality, new glottal source models that are able to characterize the acoustic features of various voice types have been proposed (Lalwani and Childers, 1991). By systematically varying the glottal source parameters and listening tests, Lalwani (1992) has reported that his source model has the potential for synthesis of speech with various voice types. However, because of the lack of a robust speech analysis procedure, the sample space (synthesized speech) of Lalwani's experiment was limited, and that made it difficult for him to reach a conclusion concerning the essential features of the various voice types.

1.2.4 Summary of Previous Research

Due to the objective measures that have become more and more popular for physiological and pathological studies of the human speech production apparatus, the appropriate and efficient speech synthesis scheme that is capable of characterizing the objective measures through an analysis–synthesis process has been studied and anticipated by speech scientists.

There are four factors that influence the use of a speech synthesizer for various applications, namely the complexity of the synthesis system, the modeling algorithm of the synthesis scheme, the quality of the synthesized speech, and the controllability of synthesis

parameters. Table 1–2 summarizes these factors for articulatory, formant, and modified LP speech synthesizers. Because of the difficulty of controlling the synthesis parameters, the articulatory synthesizer is not a good candidate for analysis–synthesis applications. The formant synthesizer has been acknowledged to be a faithful tool for psychoacoustic studies. However, due to the lack of a robust procedure for providing reliable formant information for unvoiced sounds, the sentence–based synthesis results (including voiced and unvoiced sounds) are not reliable. On the other hand, while the LP synthesizer does not model the speech production mechanism well, it is quite good at generating high quality speech for all types of phonations because of the robustness of the procedure.

Table 1–2. Summary of the advantages and disadvantages for the articulatory, formant, and modified LP synthesizer

	Articulatory synthesizer	Formant synthesizer	Modified LP synthesizer
Complexity	High	Low	Low
Modeling algorithm	models the articulatory speech production mechanism	models the acoustic speech production mechanism	models the speech signal
Synthesis quality	High (assuming synthesis parameters are well controlled)	High (assuming synthesis parameters are well controlled)	High
Controllability of synthesis parameters	Difficult	Medium for voiced Difficult for unvoiced	Simple and can be automatic

Since most psychoacoustic studies, such as the variations of vocal quality and gender conversion, are interested in acoustic aspects of features when voiced sounds are phonated,

the hybrid model that uses formant and LP synthesizers to model voiced and unvoiced sounds, respectively, becomes an attractive synthesis scheme (Olive, 1992). However, since the development of this kind of synthesis scheme is still preliminary, there is little literature regarding the use of this configuration for analysis–synthesis applications.

1.3 Research Plan

In this study, the research plan is split into two separate but related phases as shown in Figure 1–6. In phase one, the goal is to establish a complete synthesis-by-analysis system that can mimic original speech. The idea of synthesis–by–analysis is to produce synthetic speech by deriving the synthesis parameters from an original speech token. As known, the procedure of extracting the synthesis parameters is crucial for producing high quality speech. Therefore, in addition to the design of the hybrid synthesizer, the study of a robust speech analysis procedure is included in the first phase of research. This procedure, as shown in Figure 1–7, is founded on the linear prediction technique. The processes, such as formant estimating and smoothing, pitch and gain contour extraction, and glottal inverse filtering (GIF), are studied and included in order to ensure that adequate and accurate information will be provided for mimicking a voice. Note that, for voiced sounds, the waveforms that are derived from the glottal inverse filtering process are considered to be the differentiated glottal volume–velocity if the supra–glottal system is accurately represented. For convenience, this quantity will be illustrated as the estimated glottal waveform from the GIF process throughout this study.

In the second phase, based on the knowledge gained in the first phase of research, we will use the analysis–by–synthesis technique to examine the significant glottal features for various voiced types. Glottal modeling is the process of characterizing the estimated glottal waveform from the GIF by certain glottal source models, and is the essential work of this phase of research. Numerous glottal excitation models have been studied for various

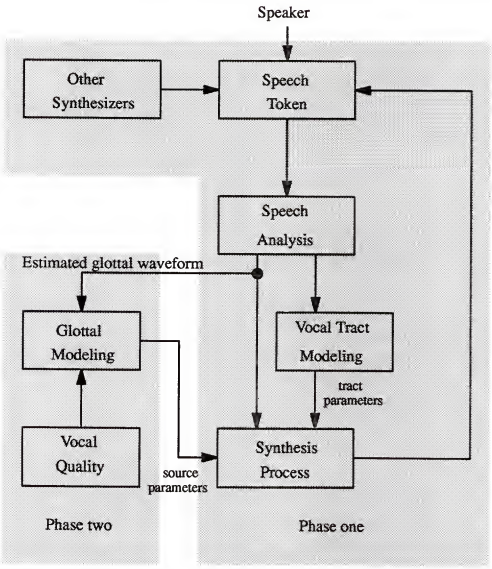


Figure 1–6. Block diagram of the research plan

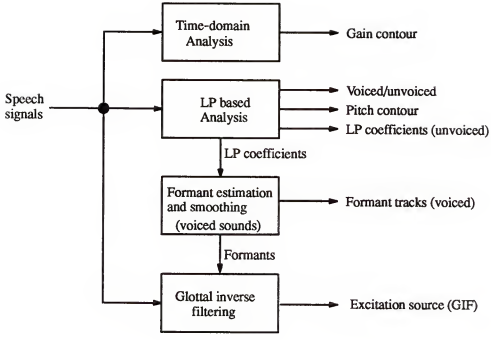


Figure 1–7. Block diagram of the analysis procedure.

kinds of speech applications (Fant et al., 1985; Fujisaki and Ljungqvist, 1986; Childers and Hu, 1994). We will investigate some of these models and implement two of them as the source models for the hybrid synthesizer.

Since we have hypothesized that vocal quality is mainly affected by the glottal excitation during the synthesis/analysis process, the vocal tract parameters will not be varied, and only the glottal source parameters will be adjusted to examine the possible causes of various voice types.

1.4 Description of Chapters

In Chapter 2, we will briefly review the basic concept of the formant and LP synthesis scheme, then introduce the configuration of the hybrid synthesis scheme. The definitions of the synthesizer parameters and the usage of the synthesis system will be presented.

For providing synthesis parameters to the hybrid synthesizer, a robust LP-based speech analysis procedure will be introduced in Chapter 3. This procedure starts from an asynchronous LP analysis and ends with glottal inverse filtering. The fundamental algorithm for each part of this analysis as well as the implementation details will be discussed.

Chapter 4 consists of two parts. In the first part, the modeling process for the glottal waveform will be presented. The estimated glottal waveform can be modeled as the superposition of two types of excitation sources: the pulse-like glottal excitation source and the aspiration noise. Each type of source will be represented by the parametric model. The significance of the modeled parameters in synthesizing various voice types will be discussed in the second part. An overview of previous research results that concern the causes of various voice types will be introduced. The procedure of using the analysis–by–synthesis technique to examine the relationship between voice type and the

modeled glottal source parameters will be addressed. An example illustrating the conversion of voice type will be illustrated at the end of this chapter.

Chapter 5 will introduce the graphic user interface for the synthesis procedure as well as the glottal source modeling process. Chapter 6 will summarize the results of this study and recommend future work.

CHAPTER 2

A FORMANT-BASED LP SYNTHESIZER

The purpose of this chapter is to design a system for synthesizing high quality speech. Based on the structure of the source-tract model, a formant-based linear predictive (LP) synthesizer is presented in this chapter. Two types of synthesis schemes constitute this hybrid synthesis system: the formant and the LP synthesizer. In Section 2.1, we review recent work concerning the two synthesis schemes. The configuration of the formant–based LP synthesizer and the synthesis parameters are addressed in Section 2.2, and Section 2.3, respectively. In Section 2.4, we will discuss the implementation details for the formant–based LP synthesizer.

2.1 Introduction

The source-tract theory that attempts to acoustically approximate the speech production process was proposed in the early 1960s (Fant, 1960). This model has been widely used not only in generating high quality speech but also in studying the acoustic aspects of speech production (Klatt and Klatt, 1990; Carlson et al., 1991; Childers and Hu, 1994; Fant, 1993; Karlsson, 1992; Carlson, 1993). A simplified schematic diagram of the source-tract model is illustrated in Figure 2–1. Theoretically, three basic functions constitute this model: 1) the source filter, G(z), 2) the vocal tract filter, V(z), and 3) the radiation filter, R(z). Since these filters are assumed to be linearly connected, the effect

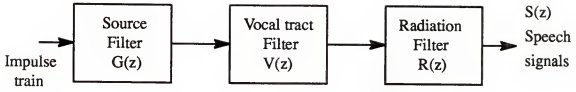


Figure 2–1. Schematic diagram of a source–tract model

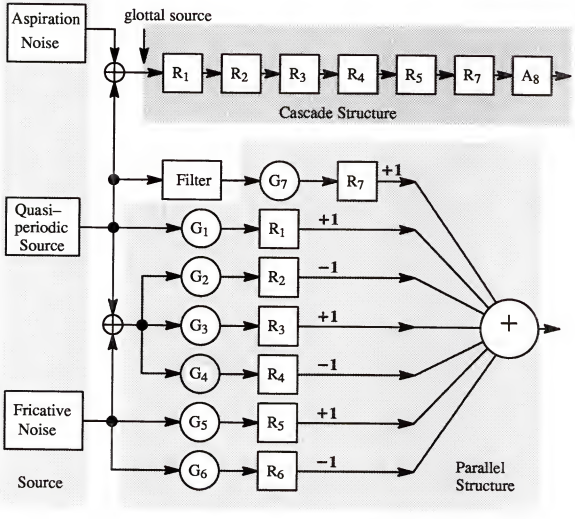


Figure 2–2. Block diagram of a cascade/parallel formant synthesizer.

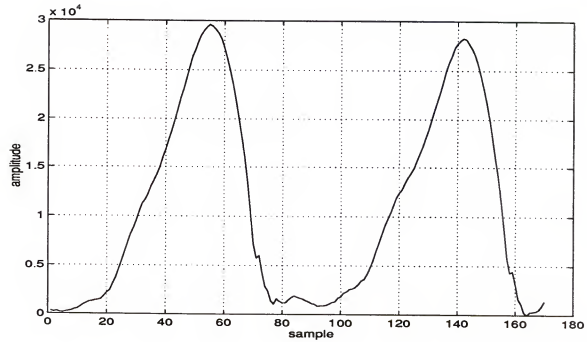
of the radiation filter is usually included in either the source filter or the vocal tract filter. Formant synthesizers and LPC synthesizers are examples of the source-tract model.

2.1.1 Formant Synthesizer

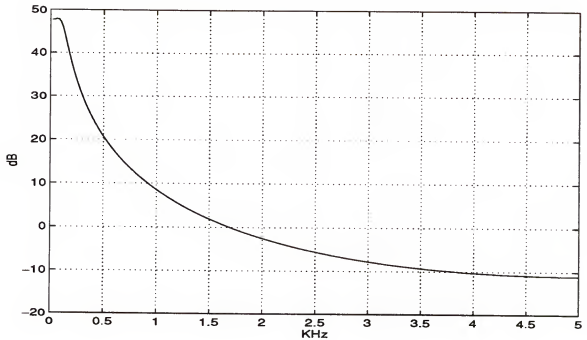
The general configuration of a formant synthesizer is shown in Figure 2–2. Three types of excitations are applied to the synthesis system: 1) the quasi-periodic source, 2) the aspiration noise, and 3) the fricative noise (Klatt, 1980; Pinto et al., 1989; Lalwani, 1991). Normally, the superposition of the quasi–periodic source and aspiration noise is called the glottal source or voicing source, by which the voiced sounds are generated. The fricative noise is employed to produce phonations, such as the unvoiced fricatives and stops. In addition to the excitation sources, two structures are generally used for simulating the vocal tract filter: 1) the cascade, and 2) the parallel structure. In these structures, the R_i block simulates the ith resonance of the vocal tract, and the A_i block simulates the ith anti–resonance of the vocal tract. The F_7 block is a first order digital filter that processes the quasi–periodic source before it is applied to the resonator, R_7. The G_i block controls the gain of each parallel branch.

2.1.1.1 Excitation source

Since the vocal folds are located below the pharynx, it is difficult to measure directly the glottal volume–velocity. Several studies have examined the characteristics of glottal flow either by direct observation of the vocal folds (laryngeal stroboscopy; high speed laryngeal cinematography) or by indirect measurement of glottal flow (electroglottography; ultra–sound glottography; and inverse filtering of speech signals) (Allen and Hollien, 1973; Childers et al., 1983; Childers and Krishnamurthy, 1985;

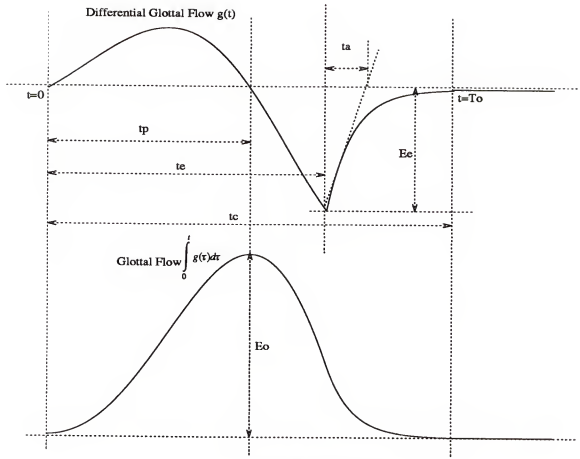


(a)



(b)

Figure 2–3. Characteristics of the glottal flow.
(a) Typical glottal waveform;
(b) Spectrum of the glottal waveform.



Glottal Parameters

$$g(t) = \begin{cases} E_e e^{\alpha t} \sin \omega_g t & 0 < t \leq t_e \\ -\dfrac{E_e}{\varepsilon t_a}\left[e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)}\right] & t_e \leq t \leq t_c \leq T \end{cases}$$

with the following restrictions

$$\int_0^T g(t) = 0 \qquad \omega_g = \frac{\pi}{t_p}$$

$$\varepsilon t_a = 1 - e^{-\varepsilon(t_c - t_e)} \qquad E_o = -\frac{E_e}{e^{\alpha t_e} \sin \omega_g t_e}$$

T = *pitch period*

Figure 2–4. The LF model: waveforms and mathematical expressions.

Hamlet, 1981; Alku, 1992). Typical waveforms and spectra of the glottal flow are shown in Figure 2–3.

The acoustic features of the glottal source can be characterized in both the time and frequency domain. Lalwani (1991) summarized that "pitch period," "glottal flow pulse width," "glottal flow skewness," "abruptness of closure," "aspiration noise," "pitch perturbation," and "amplitude perturbation" are essential time–domain features of the glottal waveform, whereas, "spectral tilt," "harmonic richness factor," and "harmonic to noise ratio" are important frequency–domain factors to characterize the glottal source.

Based on various characteristics of the glottal flow, numerous glottal models have been proposed. The "two–three pole model" proposed by Childers and Lee (1991) is an example that simulates the glottal flow in the frequency domain. This model normally adopts a second order all–pole filter to simulate the low–pass spectral behavior of the glottal source. In the case of breathy sounds, the order of the low–pass filter would be extended to three. Due to the lack of the time–domain controllability, the use of this frequency–domain source model is primitive.

By separately specifying the glottal waveform as the rising and falling segments, Fant (1979) used three parameters: 1) the glottal flow peak U_o, 2) the glottal frequency ω_g, and 3) the asymmetry factor K to model the glottal flow. Because it always produces abrupt glottal closure, Fant's model is not adequate to simulate the cases when the glottal closure is smooth.

Instead of modeling the glottal flow, Fant et al. (1985) proposed one model to simulate the differentiated glottal flow. This model is called the LF model and is illustrated in Figure 2–4. The modeled waveforms, g(t), can be specified either by the direct synthesis parameters (E_o, α, ω_g, and ε) or by the timing parameters (t_p, t_e, t_a, t_c). The parameter t_p denotes the instant of the maximum glottal flow; the parameter t_e illustrates the instant of the maximum negative differentiated glottal flow; the parameter t_a is the time constant of the exponential curve of the second segment of the LF model; the parameter

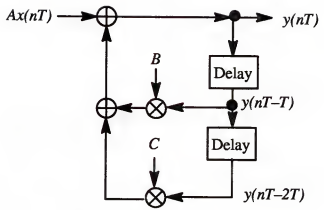
t_c is the moment when the complete glottal closure is reached. Because the LF model has one more dimension of freedom than the Fant's (1979) model, the LF model allows a control of not only the skewness of the excitation waveforms but also the manner of glottal closure. Another interesting feature of the LF model is the capability of allowing users to specify the glottal waveforms through the timing parameters, and, as is well known, the arrangement of the timing parameters is crucial for psychoacoustic studies (Childers and Ahn, 1994).

Lalwani and Childers (1991a) have added modulated aspiration noise, jitter (pitch perturbation), and shimmer (amplitude perturbation) to the LF model to form a more complete glottal model. This model characterizes the glottal flow in both time and frequency domain, and has the potential to be an excitation source for various voice types.

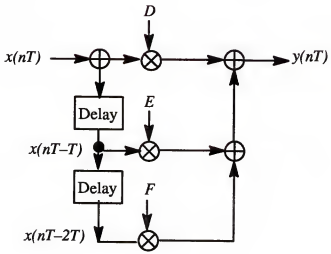
The acoustic features of the turbulent flow (fricative noise) for generating voiced fricatives, stops, and affricates have been studied by speech scientists (Stevens et al., 1992; Stevens, 1993a; Stevens, 1993b). Normally, the high–pass gaussian noise along with timing information are employed to simulate the turbulent flow in the synthesis process (Klatt, 1980; Holmes, 1983; Childers and Lee, 1991).

2.1.1.2 Vocal tract transfer function

The R_i and A_i blocks in Figure 2–2 are used to simulate the poles and zeros of the vocal tract transfer function, respectively. The digital domain implementations of these blocks are shown in Figure 2–5. The input is x(nT), and y(nT) is the output of the filter at the instant "nT," where T is the sampling period of the discrete system. The delay of the input and output signals is denoted as x(nT–iT) and y(nT–iT) respectively. According to Figure 2–5, R_i and A_k can be formulated as,



(a)



(b)

Figure 2–5. Digital realization of the second order filters. (a) Resonator; (b) Anti-resonator.

$$V_i(z) = \frac{Az}{1 - B\,z^{-1} - C\,z^{-2}} \qquad \text{For resonator i} \qquad (2-1)$$

$$A_k(z) = D + E\,z^{-1} + F\,z^{-2} \qquad \text{For anti-resonator k} \qquad (2-2)$$

$$C = e^{-2\pi bw\,T} \qquad (2-3)$$

$$B = 2\,e^{-2\pi bw\,T}\cos(2\pi f T) \qquad (2-4)$$

$$A = 1 - B - C \qquad (2-5)$$

$$F = -C/A \qquad (2-6)$$

$$E = -B/A \qquad (2-7)$$

$$D = 1.0/A \qquad (2-8)$$

where A, B, C, D, E, and F are determined by the formant (anti-formant) frequency, f, and the formant (anti-formant) bandwidth, bw, as well as the sampling time, T.

In the cascade structure, the vocal tract transfer function, V(z), is formulated as,

$$V(z) = \prod_{i=1}^{p} V_i(z) \prod_{k=1}^{q} A_k(z) \qquad (2-9)$$

where p is the total number of formants, and q is the total number of anti–formants in the cascade structure.

In the parallel structure, V(z) is formulated as,

$$V(z) = \sum_{i=1}^{p} \text{sign}[V_i(z)]\; G_i\; V_i(z) \qquad (2-10)$$

where G_i is the gain factor of each parallel branch, and sign[V_i (z)] indicates the polarity of each parallel branch. The polarity factor was suggested by Holmes (1973) in order to

match the spectrum of natural and synthesized speech. The polarized waveforms from all parallel branches are summed together to form the synthesized speech.

Based on the above general description, several formant synthesis systems have been proposed. Holmes (1973) adopted the parallel configuration to developed a formant synthesizer. In Holmes implementation, he used four parallel branches to simulate the first four formants of the vocal tract. The first three formants are implemented by the digital resonator as shown in Figure 2–5(a), and the fourth formant is accomplished by a broadband filter. A band–pass filter is connected with each branch in order to prevent the interference from one formant to the others. By using this all-parallel structure and properly controlling the synthesis parameters, Holmes et al. (1990) has generated high quality male and female speech.

In Klatt's (1980) software implementation, the formant synthesizer combined the cascade and parallel structure together. The cascade branch is used for synthesizing voiced and nasal sounds, and the parallel branch is used for generating fricative and stop phonations. By carefully designing the synthesis rules and controlling the synthesis parameters, this synthesizer is capable of generating speech with high intelligibility.

Based on Klatt's (1980) design, Lalwani (1991) developed a more flexible synthesis system. The synthesized speech can be generated by either a cascade or a parallel or a combined configuration. The connections between the excitation sources and the vocal tract all can be specified by the user. For each synthesis frame, the number of formants are flexible. Unfortunately, Lalwani did not address a way to control his flexible system. This limits the applications of the synthesizer.

2.1.2 Source Modeling for the LP Synthesizer

The theory of the LP synthesizer has been introduced in Section 1.2.2.3. For convenience, we replicate Eq.(1–4) here.

$$S(z) = \frac{1}{1 - a_k \sum_{k=1}^{p} Z^{-k}} E(z) \qquad (2-11)$$

$$V(z) = \frac{1}{1 - a_k \sum_{k=1}^{p} Z^{-k}}$$

where S(z) is the speech signal, E(z) and V(z) are thought to be the source and tract function for the LP synthesizer, respectively. A typical realization of the LP synthesizer is shown in Figure 2–6. The direct–1 configuration is employed to simulate the tract filter, V(z).

In the conventional LP synthesizer, the excitation signal, E(z), can either be an impulse train for voiced speech or be a sequence of random noise for unvoiced speech (Atal and Hanauer, 1971). These excitations will result in unnatural synthesized sounds. Attempts have been made to use more realistic waveforms instead of a simple impulse train or noise as the excitation source for generating high quality speech (Milenkovic, 1993; Childers and Hu, 1994). Since these studies are interested in not only quality but also transmission bandwidth (bit–rate), signal coding techniques are widely used to represent the excitation source. Childers and Hu (1994) have empirically found that a glottal excitation code book with 32 entries, each entry is a vector of 6th order polynomial, can give good results for voiced sounds. They also determined that two groups of 64 stochastic codebook entries are satisfactory for unvoiced sounds. The iterative analysis–by–synthesis processes are used for searching the best fit excitation codeword for each synthesis frame. Since the excitation source is coded, the above synthesis technology is known as the code excited linear predictive (CELP) synthesizer or glottal–excited linear predictive (GELP) synthesizer.

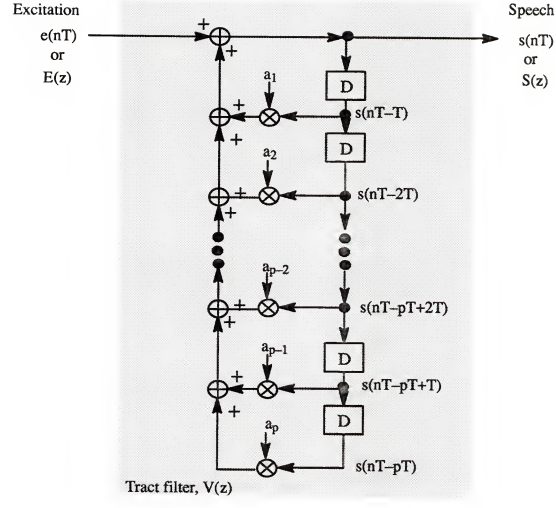


Figure 2–6. A typical realization of the LP synthesizer.

2.1.3 Time–varying Digital System

Figure 2–7 illustrates the basic concept of a digital system. By assuming that the length of the input excitation, e(n), is N, and the length of the system impulse response, v(n), is M, through the convolution process, the length of the output waveform will be N+M–1 (Strum and Kirk, 1988). For the time–invariant system, since the system impulse response is fixed, the convolution process can be done by the use of an overlap–add procedure (Strum and Kirk, 1988).

For the time–varying system, such as the LP speech synthesizer, the impulse response of the vocal tract, v(n), is updated at the beginning of each synthesis frame, thus, a more delicate process must be invoked in order to manipulate the output waveform around each frame boundary. Ignoring the time–varying effect of the vocal tract may cause amplitude discontinuity in the output waveform and generate click sounds (Lalwani, 1991).

One way to manipulate the time–varying effect of the vocal tract is the multiple–branch superposition method, as shown in Figure 2–8 (Verhelst and Nilens, 1986; Lalwani, 1991; Hu, 1993). There is one branch (with zero initial state) that takes care of the input excitation of current synthesis frame. The remaining branches (with non–zero initial state) are responsible for manipulating the residual response from previous frames. The synthesized waveform is obtained by summing the outputs of all branches together. This multiple–branch superposition method is theoretically achievable. However, since it is unrealistic to use an infinite number of branches to manipulate the time–varying effect, a threshold for the output power of each branch is used to determine the significance of the output waveform from that frame.

A more complicated method to manipulate the time–varying effect has been addressed in Hu's GELP synthesizer (1993). By assuming that the speech waveforms in many adjacent pitch periods are similar, Hu's (1993) algorithm uses two branches to handle

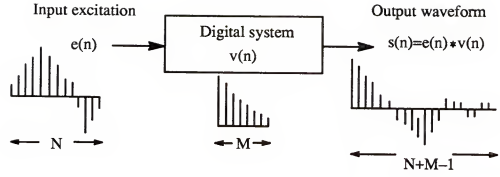


Figure 2–7. The basic concept of a digital system.

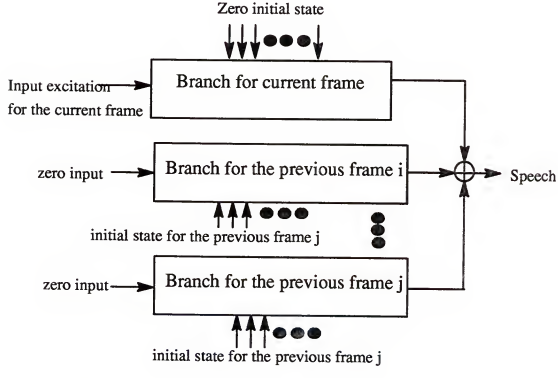


Figure 2–8. The multiple–branch superposition method.

the residual response from the previous frame. One branch with the old LP coefficients processes the residual response that is dying out, and the other branch with the current LP coefficients deals with the current input excitation. Through iterative estimate of the gain factor for the excitation, Hu's synthesis results are empirically free of click noise.

2.1.4 Hybrid Synthesis Model

In Chapter one we have already discussed the advantages of adopting a hybrid synthesis scheme to form the framework for the analysis–synthesis applications. It is quite convincing to use the formant representation to produce voiced sounds and the LP synthesis scheme to generate unvoiced sounds.

The hybrid scheme was first proposed by Olive (1992). His work used a 16th order polynomial A(z) to describe the reciprocal of the vocal tract transfer function, V(z). For unvoiced sounds, A(z) is equal to a 16th order polynomial $\hat{A}(z)$, which is derived by applying a 16th order LP analysis to the speech signal. For voiced sounds, A(z) is equal to B(z) times C(z), where B(z) represents a polynomial that relates to the formants, and C(z) is a correction polynomial that approximates the spectral difference between $\hat{A}(z)$ and B(z). By assuming that the order of polynomial $\hat{A}(z)$ and B(z) is "p" and "q," respectively, the order of polynomial C(z) will be "p–q." The algorithm for finding the correction polynomial, C(z), is to apply a LP analysis of order "p–q" to the inverse filtered signal that obtained by an inverse filtering process. The inverse filtering process is to inversely filter the speech signal by B(z).

In Olive's (1992) work, the "C(z) finding algorithm" is executed throughout the whole speech signals. In the region where the variation of the correction polynomial is bigger than a specific threshold, the region is classified as unvoiced, and thus the LP representation is adopted. Otherwise, the region is classified as voiced, and the combined

polynomial, $B(z) \times C(z)$, is used to replace $A(z)$. When multi–pulse waveform of unlimited bandwidth is employed as the input excitation, the hybrid synthesis scheme suggested by Olive (1992) has been shown to be capable of generating speech with the same quality as the modern LPC synthesizers.

2.2 Synthesis Configuration

Based on the concept of the hybrid model, our synthesis system, a 12th order formant-based LP synthesizer is illustrated in Figure 2–9. Depending upon the classification of voiced/unvoiced sounds, two categories of excitations are scaled and added at the input of vocal tract filter. They are the glottal source and the fricative noise. Both excitations can be specific waveforms, such as the estimated glottal waveform from the glottal inverse filtering, or parametric models. In generating voiced sounds, the glottal source is on, and the fricative noise is off. Otherwise, the fricative noise is chosen to excite the tract filter. A gain adjustment block is inserted between the excitations and the vocal tract filter in order to scale the input excitation and let the output speech meet the power requirement.

For both voiced and unvoiced phonations, the vocal tract filter is realized in a linear predictive configuration. Note that only one vocal tract filter is employed in this synthesis system instead of multiple branches that have been suggested by other studies to manipulate the time–varying effect of the vocal tract (Lalwani, 1991; Hu, 1993).

The vocal tract filter for the unvoiced sound is obtained from a 12th order LP analysis. On the other hand, the vocal tract filter for generating the voiced sound is not directly derived from a LP analysis but from a polynomial expansion process that multiplies six second order polynomials together, and each second order polynomial is associated with a specific set of formant frequency and bandwidth.

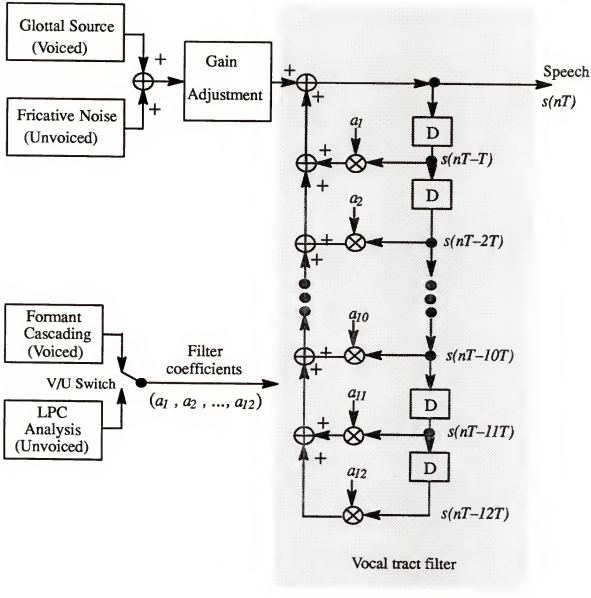


Figure 2–9. Block diagram of the 12th order formant–based LP synthesizer.

2.3 Synthesis Parameters

Before we discuss the implementation of the formant–based synthesis system, the synthesis parameters are introduced in this section. The synthesis parameters are grouped into three categories: 1) vocal tract parameters, 2) excitation parameters, and 3) control parameters. During the synthesis process, except for certain control parameters, the other parameters are updated at the beginning of every synthesis frame. The control parameters are usually kept constant for the entire synthesis process.

2.3.1 Vocal Tract Parameters

For synthesizing voiced sounds, the parameters, "f1," "f2," "f3," "f4," "f5," "f6," determine the resonant frequencies in Hz for the first six resonators of the vocal tract. These formant frequencies generally change slowly except for such situations as stop sounds. Normally, for the speech processing environment with 10 kHz as the sampling rate, the first five formants are adequate to represent the vocal tract for voiced sounds (Klatt and Klatt, 1990). However, there are exceptions where sometimes the relatively close "f4" and "f5" formants around 3.5 kHz will result in the sixth formant "f6" at the vicinity of 5 kHz (Gobl, 1988). Therefore, the first six formants are included to simulate the vocal tract. Another reason that supports the decision to select the first six formants in the formant-based LP synthesizer is from the implementation point of view, which will be discussed later. For those synthesis processes that use less than six formants to describe the vocal tract, the unspecified formant frequencies are simply set to be zero. The parameters, "b1," "b2," "b3," "b4," "b5," and "b6," specify the bandwidths of the first six formants in Hz for generating voiced sounds. The vocal tract transfer function is formed by applying the formant frequencies and their associated bandwidths to Eq.(2–1) and Eq.(2–9). Notice that this transfer function would be of order 12.

For reducing the number of variables that are interacted between the parameter specifying process and the synthesis operation, the formant parameters, "f1," "f2," "f3," "f4," "f5," "f6," "b1," "b2," "b3," "b4," "b5," and "b6," are employed to represent the LPC coefficients a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{10}, a_{11}, a_{12}, respectively, for the unvoiced sound. Since only one kind of sound (voiced or unvoiced) can be synthesized in each synthesis frame, there is no problem to adopt different definitions for the variables. The fundamental frequency "f0" is used to determine if the voiced sound is synthesized in the current frame. When "f0" is equal to zero, the unvoiced sound is produced, otherwise, the voiced sound is expected. The LPC order, 12, is chosen according to Hu's (1993) study. His results showed that the 12th order GELP (glottal excited linear predictive) synthesizer can successfully re–synthesize speech of voiced and unvoiced sounds.

2.3.2 Excitation Parameters

In addition to Lalwani's (1991) seven types of glottal sources, one modified LF source model is added to the formant-based LP synthesizer. The parameter "src_typ" is used to specify the type of glottal source that is used in the synthesis process. These glottal sources can be divided into two categories: 1) the non-parametric waveforms, and 2) the parametric models. The non-parametric waveforms such as the estimated glottal waveform from the glottal inverse filtering (GIF), are usually provided by a speech analysis procedure. The non–parametric waveforms are stored in a file sample by sample. The parameter "fil_name" is used to specify the file that stores the sampled excitation waveforms.

For the parametric source models, instead of describing the parameters that belong to any particular glottal source model, the common parameters are presented as follows. The parameter "av" specifies the power in dB for voiced sounds. Note that this definition is different from Lalwani's (1991), which is defined as the gain of the voicing source. Later

we will discuss the reason for adopting the new definition. The parameter "f0," as mentioned before, is the fundamental frequency in Hz for the current synthesis frame. The inverse of "f0" is the pitch period that denotes the duration of current frame. The parameter "jit" introduces the perturbation extent of the fundamental frequency in percent of the mean fundamental frequency. For example, if "jit" is set to 5% and the mean "f0" is 100 Hz, the maximum frequency perturbation would be $\pm$5 Hz. The parameter "shm" has a similar property, except it specifies the perturbation extent of the voiced power "av." Note that when "av" is perturbed by "shm," the parameter "av" should adopt magnitude instead of dB as the intensity unit. Both of these perturbation parameters are useful when we try to produce a segment of sustained vowels with variable "f0" and "av." This kind of sustained vowel usually sounds more natural than the one with constant "f0" and "av." The parameter "snr" is the variable that specifies the power ratio between the pulse-like glottal excitation and the aspiration noise. The aspiration noise has been found to be important for natural–sounding speech (Lalwani, 1991). In fact, the aspiration noise is not only determined by "snr" but also by other parameters, such as "amp1," "amp2," "offset," "dur," and "nfilt." Detailed explanation of these variables and the modeling process for specific glottal source models (LF model, polynomial model) are presented in Chapter 4.

The parameter "af" specifies the power in dB for unvoiced speech signals. The parameter "nfilt" controls the spectral tilt for both aspiration noise and fricative noise. The "nfilt" ranged from –1 to 0.99. The positive "nfilt" specifies a lowpass filtering process, and the negative "nfilt" indicates a high–pass filtered excitation noise.

2.3.3 Control Parameters

The parameter "sam_rat" specifies the number of speech samples that are synthesized per second. The default value of "sam_rat" is 10000. In other words, the sampling period is 0.1 ms, and thus, the frequency response of the synthesis system is up

to 5000 Hz. The parameter "frame_size" specifies the number of samples for each unvoiced frame. The default "frame_size" is 50. The parameter "g0" in dB is used for scaling up/down the final power of the synthetic signals.

Four spectral shaping parameters, "gfilt," "ufilt," "cfilt," and "ofilt," are used to enhance the capability of modifying the frequency response inside the synthesizer. In fact, the spectral shaping parameter is the coefficient of a first order digital filter. The first order digital filter can be either a finite impulse response (FIR) filter or a infinite impulse response (IIR) filter. The "gfilt" is part of the glottal source, the "ufilt" and "cfilt" are inserted between the gain adjustment block and vocal tract filter, and the "ofilt" follows the vocal tract filter. These filters work in a similar manner as the "nfilt" except for the locations. The default value for each parameters is 0, which means the filter works in a bypass manner.

The synthesis parameters can be manipulated through a graphic user interface (GUI) software, `my_fmtsyn`, which also handles the input/output interaction between the user and the synthesis system. The usage of `my_fmtsyn` along with two synthesis examples are presented in Chapter 5.

2.4 Implementation Details

By assuming that the synthesis parameters are well prepared either by a robust analysis procedure or by a trial and error experiment, the goal of the synthesis system is to make use of the parameters to generate natural–sounding speech. The following subsections will focus on the issues that are crucial for the formant-based LP synthesizer to accomplish this goal.

2.4.1 Gain Adjustment and Initial State

Two kinds of discontinuity problems affect the performance of a hybrid synthesis system. The first one is observed at the boundary when the synthesis scheme changes. The discontinuity is caused by the spectral differences between the LP representation and the formant representation. As mentioned earlier, to solve this kind of discontinuity problem, Olive (1992) proposed an algorithm to extend the order of the transfer function from the formant model to be the same order of the LPC representation. The drawback of Olive's (1992) work comes from the polynomial extension procedure, which introduces new poles and blurs the original formant representation for the vocal tract, and thus, the benefit of employing the formant scheme for the voiced sound is lost.

The other kind of discontinuity is caused by the time–varying effect of the vocal tract transfer function, which is observed at each synthesis frame boundary. The general strategy for solving this kind of discontinuity problem has been reviewed in Section 2.1.3. The concept of multiple–branch and the iterative gain control realization are widely used.

In the formant-based LP synthesizer, a different strategy, which adjusts the input gain and sets the initial state, is adopted in order to overcome the discontinuity problems just mentioned. Figure 2–10 illustrates the block diagram of this strategy, which can be thought of as a two–pass gain adjustment algorithm.

In the first pass of synthesis, the input excitation x(n) is directly applied to the vocal tract filter. The resulting output sample $\hat{s}(n)$ can be formulated as

$$\hat{s}(n) = \sum_{i=1}^{12} a_i \, \hat{s}(n - i) + x(n) \qquad 0 \leq n \leq N\text{–}1 \qquad (2\text{–}12)$$

where N is the number of samples in the current frame. The filter coefficients, (a_1, a_2, ..., a_{12}), and the initial state, ($\hat{s}(-1)$, $\hat{s}(-2)$, ..., $\hat{s}(-12)$), are specified at the beginning of each frame. Since we adopt the direct–I realization to form the vocal tract filter, by

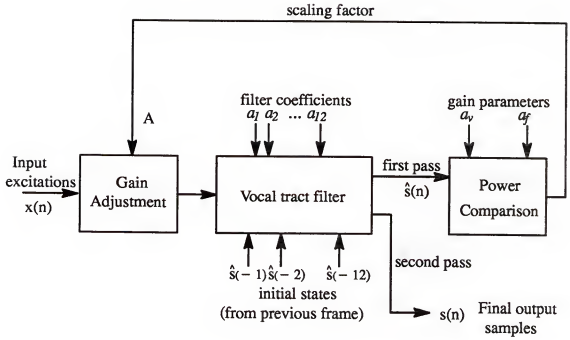


Figure 2–10. Block diagram of a gain adjustment procedure for the formant–based LP synthesizer.

assuming that the coefficients of the vocal tract filter does not drastically change from one frame to the next, the initial states of current frame can be approximated by the ending samples from the previous frame.

The output power P is

$$P = \frac{1}{N} \sum_{n=0}^{N-1} 10 \log_{10} \hat{s}(n)^2 \qquad (2\text{–}13)$$

By comparing P with the specified output power SP (av or af), a scaling factor A is defined as

$$A = \sqrt{\frac{10^{SP/10}}{10^{P/10}}} \qquad (2\text{–}14)$$

In the second pass of synthesis, the same initial states are employed, however, the input excitation, x(n), is pre–scaled by A. The final output sample, $\hat{s}(n)$, is

$$\hat{s}(n) = \sum_{i=1}^{12} a_i \, \hat{s}(n - i) + A \, x(n) \qquad (2\text{–}15)$$

Note that when less than six formants are specified for generating voiced sounds, the unspecified formant frequencies and bandwidths are set to zero. In this situation the vocal tract transfer function would still be a 12th order linear predictive filter, except that some of the higher order filter coefficients are zeros.

The following experiment is designed to verify the validity of the method that we used to manipulate the discontinuity problem. A speech token, "Should we chase those cowboys," spoken by a male speaker is first analyzed. The synthesis parameters, "f0," "av," "af," the first five formants (for voiced sounds), and the LPC coefficients (for unvoiced sounds), are specified at the beginning of every synthesis frame. In addition to the synthesis parameters, the estimated glottal waveform from the GIF is applied to the synthesizer as a non–parametric excitation source for mimicking the original speech. The

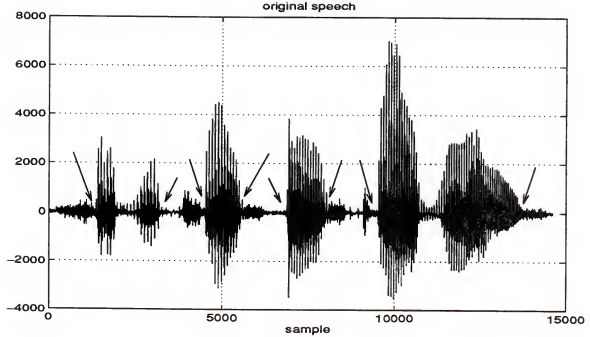
waveforms of the original speech and the synthetic one are shown in Figure 2–11. The arrow marks show the boundaries of voiced/unvoiced sounds. We observe that the two speech waveforms are similar to each other, and no abrupt transitions are shown around the arrow marks and elsewhere in the synthesized speech. An informal listening test also showed that the original and the synthetic speech are indistinguishable. The results of this experiment strongly support the synthesis algorithm just addressed.

2.4.2 Direct–I and Cascade Realization

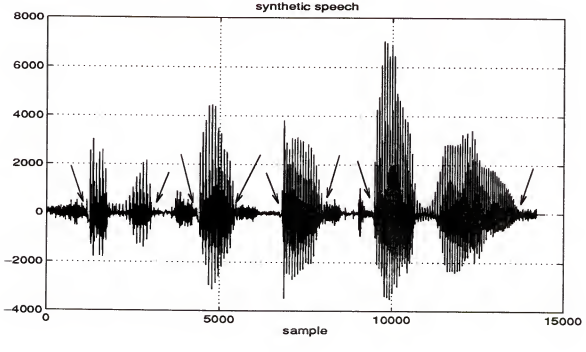
Prior to the discussion of the differences between the realizations of the vocal tract filter, we presumed that the vocal tract transfer function for either voiced sounds or unvoiced sounds was stable.

As known, the direct–I realization for the tract function is widely used in the LP synthesizer because the filter coefficients are the same as the synthesis parameters, i.e. the linear predictive coefficients. However, this is not the case for the formant synthesizer. In the formant synthesizer, the vocal tract transfer function, V(z), is usually realized by a cascade or parallel combination of a set of second order filters. This results not only from the features of the synthesis parameters (formants) but also from the consideration of the coefficient sensitivity problem (Strum and Kirk, 1988). Nevertheless, in the formant-based LP synthesizer, the direct–I realization is used to simulate the vocal tract in both synthesis schemes. The reason for this decision originates from the strategy of controlling the discontinuity problems that have been discussed in the previous subsection. The purpose of the following paragraphs is to illustrate that the direct–I realization in the formant–based LP synthesizer does not degrade the synthesis performance for the voiced regions.

Our synthesis environment is a SUN 4 system that employs 64 bits of memory to process data. These 64 bits include 1 sign bit, 11 exponent bits, and 52 fraction bits; so the



(a)



(b)

Figure 2–11. Waveforms of sentence "Should we chase those cowboys."
(a) The original speech; (b) The synthesized speech.
The arrows indicate the boundaries between voiced
and unvoiced sounds.

data of double precision format can range from $+10^{308}$ to -10^{308} with 16 significant decimal figures accuracy. In the formant–based LP synthesizer, all the calculation and storage processes are executed under the double precision format.

The following experiment is designed to examine the coefficient sensitivity problem. By applying the same excitation waveforms to three different vocal tract filters, Table 2–1 summarizes the output difference that resulted from different vocal tract realizations. In this table, the speech token 1, produced by a conventional cascade formant synthesis process, is employed as the reference. Speech token 2 is generated by the same vocal tract parameters (formant frequencies and bandwidths) as speech token 1 but realized by the direct–1 structure. Speech token 3 is synthesized in a similar manner as speech token 1 but with a 5Hz variation in its first formant frequency. The distance measure, D, is defined as

$$D = \sqrt{\frac{1}{N}\sum_{n=1}^{N}[r(n) - s(i,n)]^2} \qquad (2\text{–}16)$$

where r(n) is the nth sample of the reference token, s(i,n) is the nth sample of the speech token i, and N is the total number of samples.

It appears that the distance between the reference speech token and the speech token 2 is almost zero ($\sim 10^{-22}$), and the distance between the reference speech token and the speech token is relatively large ($\sim 10^{2}$). As well known, a 5 Hz first formant drift is insignificant in the formant synthesis, thus, the performance degradation due to the direct–1 realization is negligible and the coefficient sensitivity problem is not a problem in the double precision environment.

Table 2–1. Summary of the effects that are caused by different realization methods and different vocal tract transfer functions.

	Speech token 1 (reference token) FQ / BW	Speech token 2 FQ / BW	Speech token 3 FQ / BW
First formant	450 / 45	450 / 45	455 / 45
Second formant	1450 / 145	1450 / 145	1450 / 145
Third formant	2450 / 245	2450 / 245	2450 / 245
Forth formant	3300 / 330	3300 / 330	3300 / 330
Fifth formant	3750 / 375	3750 / 375	3750 / 375
Sixth formant	4700 / 470	4700 / 470	4700 / 470
Realization method	Cascade	Direct–I	Cascade
Distance measure	No distance measure	$\sim 10^{-22}$	$\sim 10^{2}$

Note: FQ indicates the formant frequency and BW represents the formant bandwidth.

2.4.3 Roundoff Error and Stability

Stability may be defined in the manner of the bounded input and bounded output behavior of a system (Strum and Kirk, 1988). When a bounded input is applied to a stable system, the output will be bounded. In the discrete linear causal system, the bounded input and bounded output stability can be examined by inspecting the Z–domain poles of the transfer function. For a causal system, if all the poles are located inside the unit circle, the system is stable, otherwise, the system is unstable.

In the previous section we have assumed that the vocal tract transfer function is stable. Is this assumption always true? In generating voiced sounds, yes, the system is always stable. Because in such a situation, the vocal tract transfer function is obtained by cascading a sequence of stable second order subsystems, and the cascading process will not introduce unstable poles. The second order subsystem is stable because its poles are calculated from real formants and they are located inside the unit circle.

For unvoiced sounds, the stability of the vocal tract filter needs to be explored further. This unvoiced vocal tract filter that is obtained from a LP analysis (autocorrelation method) is always stable if the filter coefficients are manipulated in a double precision format. However, using the double precision format to describe the LP coefficients is not practical in the formant–based LP synthesizer, a roundoff representation is demanded. The question is what order of roundoff error is acceptable? Since not only the LP techniques but also the acoustic features of the unvoiced speech signals will affect the characteristics of the LP coefficients, the analytical solution to the above question is still unknown.

The following experiment is designed to empirically investigate the instabilities that are caused by rounding the LP coefficients. Five hundred eighty two sets of 12th order LP coefficients with double precision formats are employed as the data base for the experiment. These LP coefficients are obtained by applying the LPC analysis to the unvoiced regions of speech tokens which are produced by three male speakers and one female speaker. The original transfer functions are free from instability because the LPC analysis is based on the autocorrelation method (Kay, 1988). By rounding the original LP coefficients to sets of new coefficients with 1, 2, 3, 4, or 5 decimal figures and examining the poles of the new transfer functions, Table 2–2 illustrates the number of instabilities for each case of truncation. The results indicate that the new transfer functions which are rounded to three decimal figures still preserve the stable property. For safety, in the formant–based LP synthesizer, we will round the LP coefficients to four decimal figures.

Table 2–2. Summary of the number of instabilities when the LP coefficients are rounded.

Number of decimal figures	16	1	2	3	4	5
Number of instabilities	0	131	3	0	0	0

Note: Totally 582 sets of LP coefficients form the data base.

2.5 Summary

In this chapter we have introduced the configuration and parameters for the formant–based LP synthesizer. The implementation details for the synthesis system were discussed and evaluated through experiments. The two–pass synthesis strategy has been determined to be effective in solving the discontinuity problem. There is no click noise around boundaries between the voiced and unvoiced sounds. The direct–1 configuration performs as good as the cascade configuration when the sensitivity problem of filter's coefficients is considered. The re–synthesis process can generate indistinguishable speech when the estimated glottal waveform from the GIF process is used as input excitation. Generally, while the amount of synthesis parameters and the complexity of the formant–based LP synthesizer are drastically reduced compared to Lalwani's flexible formant synthesizer, the synthesis results from our current work are satisfactory. In the next chapter we will focus on the analysis procedure that provides proper values for the synthesis parameters.

CHAPTER 3
SPEECH ANALYSIS

3.1. Introduction

As mentioned earlier, the control of the synthesis parameters is one of the key factors for the formant–based LP synthesizer to produce natural speech. A set of well–controlled time–varying parameters can be obtained by analyzing the speech signal. The purpose of this chapter is to present a robust procedure that analyzes the speech signal and estimates sufficient information for the formant–based LP synthesizer.

Information concerning the following items is needed by the formant–based LP synthesizer.

1. Voiced/unvoiced classification: The voiced/unvoiced classification determines which synthesis scheme (formant or linear prediction) is adopted for each synthesis frame. It also specifies which excitation source (periodic pulse–like waveform or random noise) is used for each frame.
2. Pitch period and glottal closure instant (GCI): The pitch period characterizes not only the intonation but also the vocal quality of a segment of voiced sounds. The GCI is used to determine the pitch period and provide synchronous timing information for the synthesis process.
3. Power contour: Since the human aural system is sensitive to the intensity of speech, the power contours (av and af) are needed to control the intensity of synthesized speech.

4. Vocal tract parameters: The formants (frequencies and bandwidths) and LP coefficients are used to describe the vocal tract for voiced and unvoiced sounds, respectively. These vocal tract parameters characterize not only the acoustic characteristics for each segment of a phoneme but also the personal features for an individual speaker. Note that, since the movements of the vocal tract (articulators) are slow compared to the glottis, the transitions of the vocal tract parameters are smooth.
5. Excitation source: A well–modeled excitation source has proven to be important for a high quality synthesis process. But before modeling, an estimated excitation waveform should first be accurately extracted.

3.2. LP–based Analysis

Numerous algorithms have been developed to extract acoustic features via analysis of the speech signal (Rabiner and Schafer, 1978; Lee, 1992). The LP analysis is one of these algorithms, and it is widely accepted as the basis for many practical and theoretical speech studies (Deller et. al., 1993). Since most of the synthesis parameters used by the formant–based LP synthesizer can be derived via an LP analysis, a robust LP–based analysis procedure is recommended in this research. Another reason for using LP analysis is its close linkage with the synthesis system addressed in Chapter 2. In this section we will first review the mathematic background of LP analysis, then introduce the procedure that extracts the time–varying synthesis parameters from the speech signal.

3.2.1 Linear Prediction Technique

Oppenheim and Schafer (1989) have proven that any causal, rational system, S(z), can be decomposed as

$$S(z) = A_0 A_{min}(z) A_{ap}(z) \qquad (3\text{–}1)$$

where A_0 is a constant, $A_{min}(z)$ is the minimum phase component (all the poles and zeros are inside of unit circle), and $A_{ap}(z)$ is an all–pass filter (for phase adjustment). Since the speech signal is a causal and rational system, we can decompose S(z) in this manner. Further, since phase information has virtually no effect on speech perception (Deller et al., 1993), the minimum phase component, $A_{min}(z)$, becomes important in representing the speech signal.

The linear predictive (LP) or autoregressive (AR) model is a minimum phase representation for the speech signal, if 1) the order of the LP model is correct, and 2) the input excitation waveform is orthogonal to the speech signal (Deller et al., 1993). From a practical point of view, since the orthogonality condition is generally true for unvoiced sounds as well as for low pitch voiced sounds (Deller et al., 1993; Olive, 1992), and the rule for selecting the order of the LP model for the speech signal was proposed (Markel and Gray, 1976), the LP model is widely adopted to represent the speech signal.

LP analysis is the process that calculates the LP model coefficients. This process, as shown in Figure 3–1, estimates the coefficients of the prediction filter, $\hat{A}(z)$, by minimizing the average squared prediction error, $E[\hat{e}^2(n)]$, where $\hat{e}(n)$ is difference between the real speech s(n) and the predictive output $\hat{s}(n)$, and E[x] is the operation that calculates the expected value for the variable x.

The prediction filter, $\hat{A}(z)$, and the predictive output, $\hat{s}(n)$, are formulated as

$$\hat{A}(z) = \sum_{i=1}^{p} \hat{a}\ (i)\ z^{-i} \qquad (3\text{–}2)$$

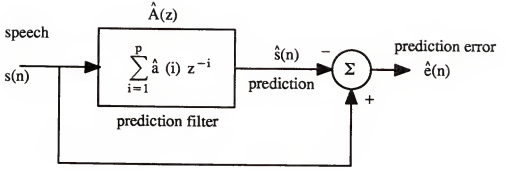


Figure 3–1. The model for estimating the prediction filter $\hat{A}(z)$.

$$\hat{s}(n) = \sum_{i=1}^{p} \hat{a}(i)\, s(n-i) \tag{3–3}$$

where $\hat{a}(i)$s are coefficients of the filter, s(n) is the speech signal, and p is the order of the LP model.

Supposed that the speech signal is wide sense stationary (WSS), the coefficients of the prediction filter, $\hat{A}(z)$, can be solved by the following Yule–Walker equation.

$$\begin{bmatrix} \Phi ss(1,1) & \Phi ss(1,2) & \Phi ss(1,3) & \cdots & \Phi ss(1,p) \\ \Phi ss(2,1) & \Phi ss(2,2) & \Phi ss(2,3) & \cdots & \Phi ss(2,p) \\ \Phi ss(3,1) & \Phi ss(3,2) & \Phi ss(3,3) & \cdots & \Phi ss(3,p) \\ \vdots & \vdots & \vdots & & \vdots \\ \Phi ss(p,1) & \Phi ss(p,2) & \Phi ss(p,3) & \cdots & \Phi ss(p,p) \end{bmatrix} \times \begin{bmatrix} \hat{a}(1) \\ \hat{a}(2) \\ \hat{a}(3) \\ \vdots \\ \hat{a}(p) \end{bmatrix} = \begin{bmatrix} \Phi ss(1,0) \\ \Phi ss(2,0) \\ \Phi ss(3,0) \\ \vdots \\ \Phi ss(p,0) \end{bmatrix} \tag{3–4}$$

$$\vec{\Phi}_{ss} \times \vec{a} = \vec{\Phi}_{ss} \tag{3–5}$$

$$\Phi ss(i,k) = E[\, s(n-i)s(n-k)\,] \qquad \text{for } 0 \le i \le p,\ 0 \le k \le p \tag{3–6}$$

where $\Phi ss(i,k)$ is defined to be the correlation function. Eq.(3–5) is the matrix–vector representation of Eq.(3–4), where $\vec{\Phi}_{ss}$ is named as the correlation matrix.

As known, the speech signal is usually stationary only for a short period, meaning that they are not WSS. Therefore, Eq.(3–6) needs to be modified for the real speech signal. Two short–term LP analysis methods are well–known and widely employed: 1) the autocorrelation method, and 2) the covariance method (Rabiner and Schafer, 1978; Kay, 1988; Marple, 1987). Table 3–1 summarizes the features for these two methods.

Table 3–1. Summary of two short–term LP analysis methods.

	Autocorrelation	Covariance
Windowing the speech waveform	Yes	No
Mean–squared error	$E = \sum_{m=0}^{N+p-1} \hat{e}^2(m)$	$E = \sum_{m=0}^{N-1} \hat{e}^2(m)$
$\Phi ss(i,k)$	$\phi_{ss}(i,k) = \sum_{m=0}^{N-1-i-k} s(m-i)s(m-k)$ $\Phi ss(i,k)=\Phi ss(k,i)$	$\Phi_{ss}(i,k) = \sum_{m=0}^{N-1} s(m-i)s(m-k)$
Correlation matrix	Toeplitz matrix	Symmetric but not Toeplitz matrix
Algorithm for solving Eq.(3–4)	Levinson–Durbin recursion	Cholesky decomposition
Stability of prediction filter	Always stable	Not guaranteed stable

Note. N is the number of the samples being analyzed.
p is the LP model order, and $1 \leq i \leq p$, $0 \leq k \leq p$.

3.2.2 Asynchronous LP Analysis

The block diagram of the analysis procedure is depicted in Figure 3–2. This procedure is a two–phase LP analysis. The first block is to pre–process the speech signal. A 20 ms silent segment (30 dB below the peak amplitude of input signal) is added at the beginning and ending of the input signal. These silent segments can assist in finding and correcting possible mistakes, such as the discontinuity of formant tracks and the mis–classification of voiced/unvoiced sounds. The resultant speech signal is normalized according to a specific peak value. The normalized waveform is then filtered by a zero–phase filter, H(z), which is capable of removing the D.C. component of the speech signal.

$$H(z) = \frac{1 - z^{-1}}{1 - 0.99z^{-1}} \qquad (3\text{–}7)$$

In the fixed frame LP analysis, one segment of speech signals are analyzed by a 12th order covariance LP method. The segment length is 250 samples with 50 samples overlapped. No pre–emphasis and windowing effect are applied to the segmented speech signal. Two types of information are determined for each segment of the speech signal: 1) the LP coefficients, and 2) the "prediction error waveform," $\hat{e}(n)$. Two examples of the prediction error waveform are shown in Figure 3–3. Significant waveshape and amplitude differences between the voiced and unvoiced sounds are observed.

Since there are 50 overlapped samples between adjacent analysis frames, the prediction error in the overlapped region is contributed by two analysis frames, e.g. frame M and frame (M+1). The prediction error, $\hat{e}(n)$, in the overlapped region is defined as,

$$\hat{e}(n) = \frac{51 - n}{51} \times \hat{e}_M(n) + \frac{n}{51} \times \hat{e}_{M+1}(n) \qquad 1 \leq n \leq 50 \qquad (3\text{–}8)$$

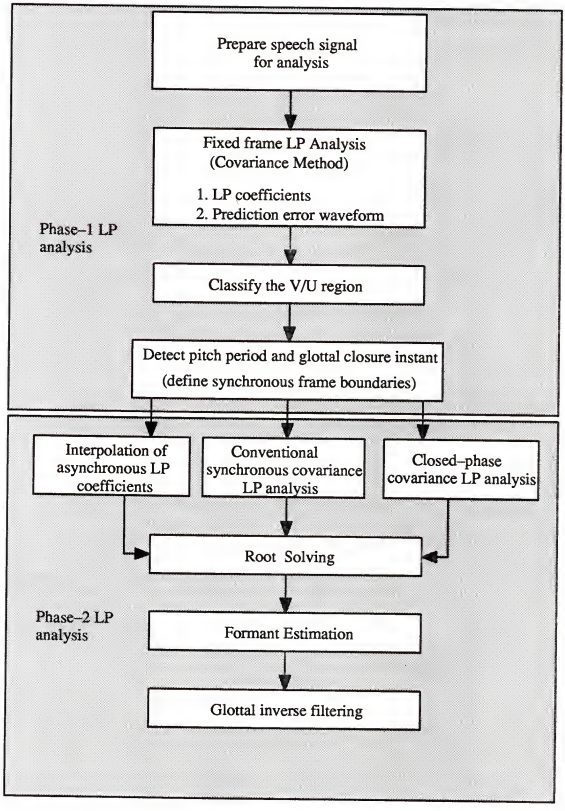
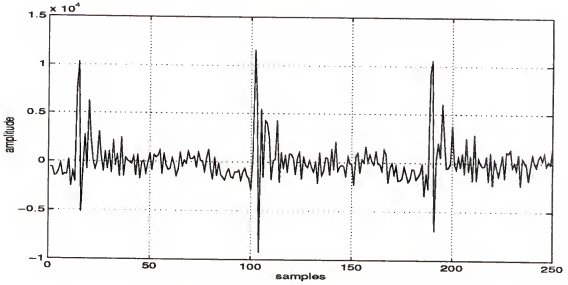
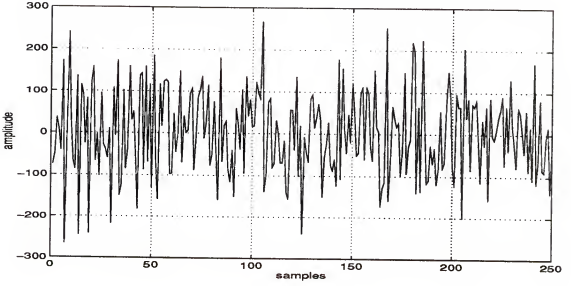


Figure 3–2. Block diagram of the analysis procedure.



(a)



(b)

Figure 3–3. Waveforms of the prediction error.
(a) For a typical voiced sound;
(b) For a typical unvoiced sound.

where n is the relative position in the overlapped region, $\hat{e}_M(n)$ and $\hat{e}_{M+1}(n)$ is the prediction error associated with the Mth and (M+1)th frame, respectively.

The following subsections will describe the other functions in Figure 3–2.

3.2.3 Voiced/Unvoiced Classification

Numerous studies have addressed the development of algorithms for classifying the speech signal into various categories, such as voiced, unvoiced, mixed, nasal, and silent (Atal and Rabiner, 1976; Childers et al., 1989a; Lee, 1992). The two–channel (speech and EGG) algorithm is a reliable solution (Childers and Krishnamurthy, 1985; Krishnamurthy and Childers, 1986; Childers et al., 1989a). However, owing to the use of the EGG signal is not practical in most situations. Other studies have tried to use only acoustic features, such as energy, zero crossing rate, level crossing rate, and spectral distribution to reach a satisfactory classification result (Hahn, 1989; Lee, 1992). In these studies, acoustic features were manipulated by signal processing techniques, such as pattern recognition, neural network, and decision tree methods in order to achieve a statistical decision. However, the results are not perfect because factors, such as the setting of thresholds, the training process, and the data base, will affect the performance of the classification algorithm.

For our application, only a two–way (V/U) decision is needed by the formant–based LP synthesizer. Thus, a relatively simple classification algorithm is developed. In Figure 3–3 we observe the significant difference between the prediction error for the voiced and unvoiced sounds, the energy of the prediction error and the first reflection coefficient are used to classify the signal as voiced. The first reflection coefficient, r_1, is

$$r_1 = \frac{R_{ss}(1)}{R_{ss}(0)} \tag{3–9}$$

$$R_{ss}(0) = \sum_{n=1}^{N} s(n) \times s(n) \qquad (3\text{–}10)$$

$$R_{ss}(1) = \sum_{n=1}^{N-1} s(n) \times s(n+1) \qquad (3\text{–}11)$$

where N is the number of samples in one analysis frame, and s(n) is the speech sample. The decision rules are:

1. If the first reflection coefficient is greater than 0.2, and the prediction error energy is greater than twice the threshold, e.g. 10^7, then this frame is voiced.
2. If the first reflection coefficient is greater than 0.3, and the prediction error energy is greater than the threshold that is used in rule "1," and, the previous frame is also voiced, then the current frame is voiced.
3. If the above conditions are not valid, then the current frame is unvoiced.

After applying the above procedure to every analysis frame, a number string is obtained (e.g. 0011100110010...). In this string, each number is associated with one analysis frame, where "1" and "0" represents a voiced and unvoiced segment, respectively. Since in real speech, the patterns, such as 101 and 010, seldom occur, a process that corrects the 101 (010) string into 111 (000) string is used to reduce the error rate of classification.

3.2.4 Detection of Pitch Contour and Glottal Closure Instant

Strategies for deriving the pitch period or the fundamental frequency can be divided into two categories: 1) a synchronous, and 2) an asynchronous strategy. The former group of algorithms, such as the SIFT (simple inverse filtering tracking), are capable of providing accurate timing information, but are sensitive to noise (Markel, 1972). On the other hand, the latter group, such as the autocorrelation and cepstrum methods, are more reliable, but

have a synchronization problem (Rabiner and Schafer, 1978; Lee, 1992; Childers et al., 1993).

Since we are interested in modeling the glottal source in this research, and the accuracy of the timing information of the glottal source is crucial for the modeling process, a reliable algorithm that can provide accurate and synchronous pitch period information is needed. The two–pass glottal closure instant identification procedure, which is proposed by Childers and Hu (1994), has proven to be satisfactory for this necessity.

Their method makes use of the V/U classification result and the prediction error waveform, e(n), to detect the pitch period and the glottal closure instant. The detection algorithm can be divided into two stages: 1) pitch estimation, and 2) peak picking.

1. Pitch period estimation:
   a. Low–pass filter one segment of prediction error waveform, e(n). The filtered waveform is denoted as $e_{Lp}(n)$.
   b. Calculate the ceptrum–like sequence, $C_e(n)$,

$$C_e(n) = \text{IFFT}\left(\left| \text{FFT}\left(e_{Lp}(n) \right) \right| \right) \qquad 1 \le n \le N \qquad (3\text{–}12)$$

   where N is the frame size, FFT is the fast fourier transform operation, IFFT is the inverse FFT, and $|x|$ is the operation that calculates the absolute value of variable x.

   c. Search for the index "m," where $C_e(m)$ has the maximum amplitude in the subset $\{C_e(i) \mid 25 \le i \le N\}$.
   d. Search for the index "k," where $C_e(k)$ has the maximum amplitude in the subset $\{C_e(i) \mid 25 \le i \le m\text{–}25\}$.
   e. If $C_e(k) > 0.7C_e(m)$, "k" is the estimated pitch period, otherwise, "m" is the estimated pitch period.

f. If an abrupt change of pitch period is observed (compared to previous pitch periods), a low–pass filtering process is invoked to smooth the abrupt change.

2. Peak picking:

a. In each analysis frame (with 256 samples) search for the most negative peak of the $e_{Lp}(n)$ waveform.

b. Build a template as illustrated in Figure 3–4 (a). This template is formed by the waveform around the most negative peak of $e_{Lp}(n)$. The length of the template is 46 samples (including 15 samples before the peak, 30 samples after the peak, and the peak itself), and the length was suggested by Hu (1993).

c. Correlate the template with the $e_{Lp}(n)$ waveform to generate a new sequence, $C_{te}(n)$, as shown in Figure 3–4 (b).

d. The positive peaks of the $C_{te}(n)$ sequence provides the possible GCIs. The estimated pitch period from stage one can assist in correcting the erroneous peaks.

e. Adjust the position of each GCI under the criteria that no two GCIs are located within 25 samples. The choice of 25 samples as the threshold was also suggested by Hu (1993).

An interactive software, `modgci`, was implemented to inspect and correct the possible error in locating GCIs.

The power contours are calculated according to the GCI information, and they are formulated as,

$$a_v = 10\log_{10}[\frac{1}{N} \times \sum_{i=1}^{N} s(i) \times s(i)] \qquad (3\text{–}13)$$

$$a_f = 10\log_{10}[\frac{1}{N} \times \sum_{i=1}^{N} s(i) \times s(i)] \qquad (3\text{–}14)$$

where a_v is the voiced power in dB, a_f is the unvoiced power in dB, and N is the pitch period for voiced sound.

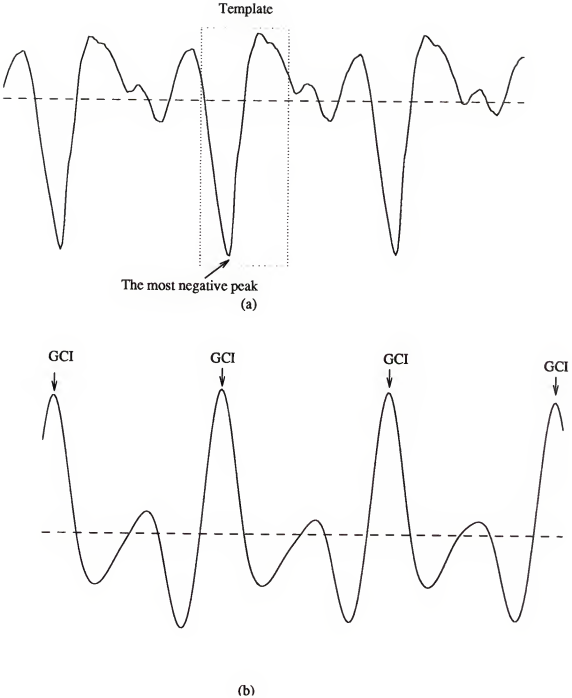


Figure 3–4. Illustration of the pitch period and GCI detection algorithm. (a) The filtered prediction error sequence, $e_{LP}(n)$; (b) The correlation output sequence, $C_{ec}(n)$.

3.2.5 Synchronous LP Coefficients

Using the pitch period and GCIs just derived, three methods have been implemented for deriving the pitch synchronous LP coefficients: 1) interpolation of the LP coefficients, 2) conventional synchronous LP analysis, and 3) closed–phase LP analysis (Olive, 1971; Larar et al., 1985; Wong et al., 1979).

1. Interpolation of LP coefficients:

   Based on Hu's (1993) work, a quadratic weighting function is employed to interpolate the LP coefficients that are obtained in Section 3.2.2. Instead of linear interpolation, the quadratic algorithm provides a sequence of synchronous LP coefficients. Since this interpolation process does not guarantee the stability for the resulting LP coefficients, a root checking process is needed to stabilize the coefficients for the filter. By applying this method, the amount of computation is considerably reduced, and the transition of the vocal tract filter is relatively smooth compared to the other two methods that will be described later. However, because of its asynchronous nature, the resultant LP filter might not exactly simulate the vocal tract. This problem becomes serious when attempts are made to decompose the speech signal into the glottal source and the vocal tract.

2. Conventional synchronous LP analysis:

   For avoiding degradation that results from the asynchronous process, a synchronous LP analysis has been developed. The covariance method is applied to the pre–emphasized voiced sounds, and the frame size is equal to the pitch period. The autocorrelation method along with a 5 msec hamming window is used to calculate the LP coefficients for unvoiced sounds. The LP orders are kept at 12 for both regions of sounds.

3. Closed–phase LP analysis:

   The closed–phase LP analysis is based on the assumption that the source–tract model is more reliable in the glottal closed region than the open region. Also, knowledge of the phase of the glottis can assist the estimation of the vocal tract parameters (Wong et

al., 1979). Both two–channel (speech and EGG signals) and single–channel (speech signal) approaches have been studied to mark the closed–phase region (Krishnamurthy and Childers, 1986; Milenkovic, 1986). Since the use of the two–channel approach is not always practical, the single–channel method is employed to calculate the closed–phase LP representation. The analysis steps are:

a. Using the GCI obtained in Section 3.2.4., five fixed–frame (with frame size 36) linear prediction covariance analyses are performed. Each frame is started from a sample around the GCI.
b. For each fixed–frame LP analysis, a prediction error sequence is obtained.
c. Based on the minimum prediction error criterion, choose the best set of LP coefficients among those five candidates.

Note that the closed–phase LP analysis is only for voiced sounds. The autocorrelation method is still used for analyzing unvoiced sounds.

3.2.6 Formant Estimation

A twelfth order LP analysis provides twelve roots, some of these roots belong to the vocal tract and others belong to the glottal source or radiation filter. The formant estimation process that assigns appropriate roots to simulate the vocal tract. The relationships between the roots and the possible formants are defined as follows,

$$FQ = \theta/\pi \times 5000 \quad Hz \tag{3–15}$$

$$BW = -\log_e(r)/\pi \times 10000 \quad Hz \tag{3–16}$$

where FQ is the possible formant frequency, BW is the possible formant bandwidth, θ is the angle of the root in radian, and r is the magnitude of the root.

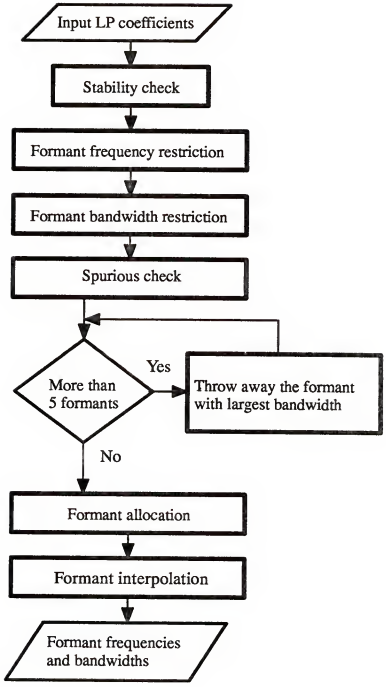


Figure 3–5. The procedure to estimate the formant tracks.

Numerous studies have estimated the formants based on various LP approaches (Markel, 1973; McCandless, 1974). It seems that the accuracy and the continuity are two main concerns in the formant estimation process.

Based on the LP coefficients that were derived in the synchronous LP analysis, a formant estimation procedure is developed and shown in Figure 3–5. The details are summarized as follows:

1. Stability check: Since the covariance method does not guarantee a stable filter, the roots that are located outside of unit circle (in Z–plane) have to be moved into inside of unit circle. The new root should have the same phase as the original root but with a magnitude that is reciprocal to the original one.
2. Formant frequency restriction: Those roots with resonant frequency less than 200 Hz or larger than 4200 Hz are not considered valid for simulating the vocal tract.
3. Formant bandwidth restriction: Those roots with a resonant bandwidth larger than 700 Hz, or a bandwidth to frequency ratio larger than 0.5 are not considered valid for simulating the vocal tract.
4. Spurious check: When the resonant frequency of two roots are close to each other (less than 200 Hz) and one of root has a bandwidth larger than 450 Hz, the root with the larger bandwidth is not a formant candidate.
5. Formant deletion: After steps 2, 3, and 4, for each individual frame, if less than six roots remain, then go to the next step. Otherwise, the one with the largest bandwidth is to be removed. The removing process will continue until only five formants remain.
6. Formant allocation: Right now each frame will have five roots or less. For the frames with exactly five roots, we skip the following process and move to the next step. For the frames with less than five roots, a procedure is used to assign the roots as the proper formants. This procedure is illustrated by an example as shown in Figure 3–6. In this example, the frames (n–3), (n–1), n, (n+3), and (n+4) have exactly five roots, so these roots are well arranged in an ascending order, meaning that the root with the smallest

formant frequency is designated as the first formant and that the root with the largest formant frequency is designated as the fifth formant. For frames (n–2), (n+1), and (n+2), there are less than five roots. The circled roots in these frames are allocated according to the continuity criteria, vacant formant slots are observed in these frames.

7. Formant interpolation: The vacant formant slots are filled by the use of a linear interpolation process. For example, the empty slot in frame (n+3) can be filled with the value, 3685.

3.2.7 Glottal Inverse Filtering

Based on the formants just derived, a 10th order transfer function, V(z), can be obtained to describe the vocal tract for voiced sounds,

$$V(z) = \prod_{i=1}^{5} V_i(z) = \frac{1}{A(z)} = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2} \ldots + a_{10} z^{-10}} \qquad (3\text{–}17)$$

where $V_i(z)$ is the transfer function for individual formant, and A(z) is the inverse of V(z).

The glottal inverse filtering is formulated as

$$E(z)=S(z)A(z) \qquad (3\text{–}18)$$

where S(z) is the Z domain representation of the speech signal, and theoretically E(z) is the differentiated glottal waveform.

3.2.8 Software Summary

A software program, `ana_inface`, was implemented in order to link all the analysis steps together. The input of this software is a segment of speech waveform and the resultant acoustic features are: 1) the fundamental frequency, 2) the power of voiced

	first formant	second formant	third formant	fourth formant	fifth formant
frame n−3	450	1300	2450	3300	3750
frame n−2		1350	2450	3350	3600
frame n−1	460	1300	2450	3300	3650
frame n	470	1350	2450	3300	3670
frame n+1	450		2450		3670
frame n+2	430	1200	2450	3300	3685
frame n+3	390	1250	2450	3350	3700
frame n+4	440	1300	2450	3340	3800

Figure 3−6. Example for formant allocation and interpolation.

sounds, 3) the power of unvoiced sounds, 4) the first five formants (for voiced sounds), and 5) the coefficients of LP polynomial (12th order). These acoustic features are stored in a file, and this file can be directly employed by the formant–based LP synthesizer as the parameter track file. The `ana_inface` also provides a file that contains the estimated glottal waveform from the glottal inverse filtering process. This file can be employed by the formant–based LP synthesizer as a non–parametric excitation source.

3.3 Experiments

Three experiments are presented to evaluate the performance of the analysis procedure. The speech signals analyzed are both real and synthesized.

3.3.1 Voiced/Unvoiced Classification

Although the V/U classification is executed in the asynchronous phase of speech analysis, it is improper to evaluate the V/U decision in an asynchronous manner, especially when one portion of a segment of speech signal is voiced and the other portion is unvoiced. Therefore, we adopt the "detection error of GCIs" as the criterion to evaluate our V/U classification algorithm. The "detection error of GCIs" occurs in one of two ways: 1) missing a GCI when it should be present, and 2) introducing a GCI when it should not be present. The decision for a erroneous detection is made by visually inspecting the speech and EGG waveforms.

Three sentences: 1) "We were away a year ago," 2) "Should we chase those cowboys," and 3) "That zany van is azure" are used to form the data base for this experiment. Each sentence was spoken by three speakers.

Based on the evaluation criterion just created, Table 3–2 summarizes the experimental results. An overall 93.2% detection rate is achieved. Compared with other

research that dealt with V/U/M/S classification (Childers et al., 1989; Lee, 1992), our result is acceptable. Two kinds of phenomena are apparent in this experiment. The first phenomenon is that most errors resulted from missing a GCI instead of introducing an excessive GCI. It seems that the classification process tends to be conservative in declaring a voiced region. The second phenomenon is that the classification algorithm performs the best for sentence 1, which is all voiced. The second and third sentences have errors that occur around the voiced consonants /b/ and /z/. Fortunately, it is not so critical to classify these phonations to be either voiced or unvoiced.

3.3.2 Pitch Detection

The performance of the pitch detection algorithm can be evaluated in the following manner:

$$PD = \frac{1}{M} \times \sum_{i=1}^{M} | d(i) | \qquad (3\text{–}19)$$

$$d(i) = \frac{f(i) - f_d(i)}{f_d(i)} \qquad (3\text{–}20)$$

where PD is the averaged shift of the fundamental frequency, M is the total number of analysis frames, f(i) is the derived fundamental frequency for frame i, and $f_d(i)$ is the true fundamental frequency for frame i.

Fifteen synthesized signals are examined in this experiment. They are different in the perturbation range of the fundamental frequency as well as the waveform. Each synthesized signal has 100 frames (pitch periods).

Based on the above definitions, Table 3–3 summarizes the distance measures for PDs. An averaged shift of 1.86% is obtained. The PD increased as the frequency perturbation increased. This result is reasonable, since the estimated pitch period from the first–pass of the pitch detection algorithm is an averaged value, and this averaged result is

Table 3–2. The detection rate of GCIs.

	total number of GCIs	number of detection errors		detection rate (%)
		miss a GCI	introduce a GCI	
sentence 1	580	23	2	95.7
sentence 2	438	41	1	90.4
sentence 3	608	40	3	92.9
total	1626	104	6	93.2

Table 3–3. The performance of the pitch detection algorithm.

Distance PD / Excitation waveform	Frequency perturbation range					
	1 %	2 %	3 %	5 %	10 %	Average
waveform 1	0.70%	0.80%	1.08%	1.64%	2.97%	1.44%
waveform 2	0.72%	0.97%	1.38%	2.31%	4.57%	1.99%
waveform 3	0.63%	0.83%	1.11%	1.60%	6.54%	2.14%
Average	0.68%	0.87%	1.19%	1.85%	4.69%	1.86%

not so meaningful when the perturbation of the fundamental frequency becomes a large value. Table 3–3 also shows that the waveform difference is not a significant factor in affecting the performance of the pitch detection algorithm. This result illustrates that the pitch detection algorithm is effective for various speech tokens.

3.3.3 Formant Estimation

This experiment is designed to evaluate the performance of the formant estimation algorithm in accordance with three pitch synchronous LP analysis procedures: 1) quadratic interpolation of asynchronous LP coefficients (QI method), 2) conventional synchronous covariance LP analysis (SC method), and 3) closed–phase LP analysis (CP method).

The measures, FD(k) (BD(k)), are used to illustrate the distance between the true and estimated frequency (bandwidth) for the kth formant, and can be formulated as

$$FD\ (k) = \frac{1}{M} \times \sum_{i=1}^{M} \mid fq_e(i,k) - fq_d(i,k) \mid \qquad 1 \leq k \leq 5 \qquad (3\text{–}21)$$

$$BD\ (k) = \frac{1}{M} \times \sum_{i=1}^{M} \mid bw_e(i,k) - bw_d(i,k) \mid \qquad 1 \leq k \leq 5 \qquad (3\text{–}22)$$

where M is the total number of frames, $fq_d(i,k)$ is the kth true formant frequency for frame i, and $fq_e(i,k)$ is the kth estimated formant frequency for frame i. Similar definitions are applied to the bandwidth components, $bw_d(i,k)$ and $bw_e(i,k)$.

Nine synthesized signals constitute the data base for this experiment. They all have the same f0 contour. Three sets of formants as well as excitation waveforms differentiate the synthesized signals. Each synthesized signal is a sequence of a vowel–like sound with 10000 samples.

Table 3–4 illustrates the performance of the formant estimation algorithm for three excitation waveforms W1, W2, and W3. We observe that the CP method can provide the exact formants when W1 or W2 is used as the excitation source. When W3 is used as the

excitation source, the CP method is still the best even a 63 Hz error occurs in the first formant. Such error can be explained by the short duration of the closed phase of the glottal waveform. Table 3–5 illustrates the performance of the formant estimation algorithm for three phonemes G1, G2, and G3. We observe that the CP method is the best for all three phonemes.

The overall performance of the formant estimation algorithm can also be assessed by comparing the true excitation waveform with the estimated excitation waveform from the glottal inverse filtering process. Figure 3–7 gives an example of the true excitation waveform as well as the estimated excitation waveforms from the three formant estimation algorithms. Based on a visual evaluation of that figure, we conclude that the CP method is the best.

3.4 Summary

In this chapter we have introduced a two–phase LP–based analysis that extracts acoustic features from speech tokens. The two–phase analysis includes the classification of voiced and unvoiced sounds, the identification of the glottal closure instant, and the estimation of formants. For the V/U classification, an overall 93.2% accuracy for detecting a true glottal closure instant is achieved. This two–way classification result is acceptable compared to other research (Hahn, 1989; Lee, 1992). The pitch detection algorithm, which provides a less than 2% averaged drift in tracking the fundamental frequency, is robust for various kinds of speech signals. The effectiveness of the three pitch synchronous LP analysis methods, which are used to estimate the first five formants, has been summarized by illustrating the numerical distance between the true and calculated formants. The similarity between the true excitation waveform and the estimated glottal waveform from the GIF was also used to assess the performance of the formant extraction procedure. For synthetic speech, when the glottal excitation waveform has longer closed–phase (above

30% of the pitch period), the closed–phase LP analysis (CP method) provides the exact formants, and the resulting waveform from the inverse filtering is the same as the true excitation waveform. The software, `ana_inface`, combines the analysis procedures together and provides: 1) the fundamental frequency, 2) the power of voiced sounds, 3) the power of unvoiced sounds, 4) the first five formants (for voiced sounds), and 5) the coefficients of LP polynomial to the formant–based LP synthesizer. The `ana_inface` also provides a file that contains the estimated glottal waveform from the glottal inverse filtering process. Based on this procedure and the formant–based LP synthesis system, in the next chapter, a complete synthesis/analysis process will be designed to study the relationships between vocal quality and the acoustic features of the glottal source.

Table 3–4. The performance of the formant estimation algorithm with respect to various excitation waveforms.

Method		FD(1) (BD(1))	FD(2) (BD(2))	FD(3) (BD(3))	FD(4) (BD(4))	FD(5) (BD(5))
W1	QI	38.5 (21.0)	24.7 (47.7)	29.0 (26.3)	15.0 (12.5)	11.0 (47.0)
	SC	12.0 (22.3)	6.0 (13.6)	19.0 (9.0)	10.3 (9.0)	13.3 (38.0)
	CP	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
W2	QI	14.5 (62.0)	13.0 (34.0)	33.7 (30.0)	15.5 (16.5)	8.7 (51.3)
	SC	15.0 (36.7)	11.3 (23.6)	12.0 (29.0)	20.0 (6.0)	8.7 (21.7)
	CP	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
W3	QI	100.5 (93.0)	45.5 (121.5)	119.3 (201.0)	53.0 (171.5)	266.0 (143.5)
	SC	44.0 (94.6)	12.3 (32.6)	48.0 (85.3)	73.0 (27.0)	33.3 (89.6)
	CP	63.0 (23.6)	6.0 (5.0)	10.6 (40.3)	36.0 (7.6)	8.6 (4.6)

Note: W1, W2, W3 indicate the group of synthesized signals associated with particular glottal excitation waveform.
W1 has LF parameters t_p=45, t_e=60, t_a=1, and t_c=65;
W2 has LF parameters t_p=20, t_e=25, t_a=1, and t_c=35;
W3 has LF parameters t_p=50, t_e=80, t_a=8, and t_c=100.

Table 3–5. The performance of the formant estimation algorithm with respect to various phoneme groups.

Method		FD(1) (BD(1))	FD(2) (BD(2))	FD(3) (BD(3))	FD(4) (BD(4))	FD(5) (BD(5))
G1	QI	20.0 (34.0)	20.6 (26.0)	70.0 (57.3)	47.0 (44.6)	140.6 (124.3)
	SC	11.0 (6.0)	4.3 (7.6)	29.6 (32.3)	28.0 (32.0)	29.3 (40.0)
	CP	7.3 (0.3)	1.0 (2.0)	2.3 (7.0)	1.0 (5.0)	4.6 (3.3)
G2	QI	N/A	36.3 (98.6)	75.0 (78.0)	N/A	43.3 (42.3)
	SC	11.6 (11.6)	9.3 (23.0)	31.0 (69.3)	71.0 (0)	10.6 (19.0)
	CP	7.6 (5.6)	3.0 (1.3)	6.0 (29.6)	33.6 (0)	1.3 (1.3)
G3	QI	82.3 (83.3)	16.5 (57.0)	37.0 (122.0)	8.6 (89)	19.5 (41.0)
	SC	48.3 (136.0)	16 (39.3)	18.3 (21.6)	4.3 (10.0)	15.3 (90.3)
	CP	48.0 (17.6)	0 (0)	2.3 (3.7)	1.3 (2.6)	2.6 (0)

Note: 1. G1, G2, G3 indicate the group of synthesized signals associated with particular formants.
G1 has f1=316, f2=2115, f3=2860, f4=3378, f5=3604, b1=40, b2=95, b3=246, b4=180, and b5=265;
G2 has f1=312, f2=833, f3=2443, f4=2738, f5=3500, b1=59, b2=93, b3=334, b4=500, and b5=242;
G3 has f1=529, f2=842, f3=2343, f4=3150, f5=4150, b1=224, b2=205, b3=144, b4=133, and b5=500.
2. N/A means the formants are not available.

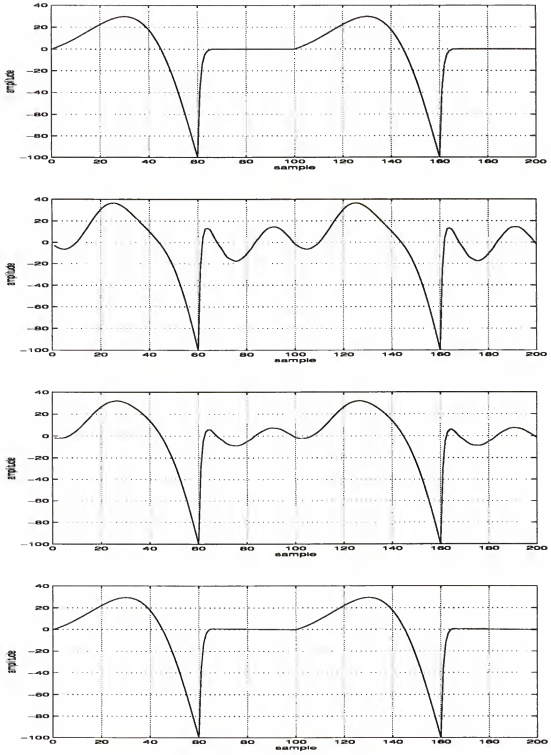


Figure 3–7. Two periods of excitation waveform.
(a) Original (true) waveform;
(b) Waveform derived by the QI method;
(c) Waveform derived by the SC method;
(d) Waveform derived by the CP method.

CHAPTER 4

GLOTTAL SOURCE MODELING AND VOCAL QUALITY

The purpose of this chapter is to use the formant–based LP synthesizer to produce sounds with various vocal qualities, such as modal, breathy, vocal fry, falsetto, harsh, and whisper (Laver, 1980; Laver and Hanson, 1981). A voice conversion approach is used to explore the essential glottal features for various types of voices. By assuming that a segment of a speech signal can be accurately decomposed into a glottal waveform and a vocal tract transfer function via the formant extraction and the glottal inverse filtering (GIF) processes (Childers and Ahn, 1994; Gobl, 1988; Fujisaki and Ljungqvist, 1986; Fujisaki and Ljungqvist, 1987; Alku, 1992), the key concept of the voice conversion is to reproduce the vocal tract component, but vary the glottal source for generating sounds with a different vocal quality. The variations of the glottal waveform can be achieved by properly modeling the glottal source and systematically varying the model parameters.

4.1 Glottal Source Modeling

Glottal source modeling is an interesting topic for applications such as high quality speech synthesis, speech communication, and psychoacoustic studies of vocal quality, because an effective glottal source model that makes use of the essential acoustic features of the glottal source can assist speech scientists to efficiently resolve the problems concerned.

According to previous studies, three types of features characterize the glottal source: 1) the low frequency waveform, 2) the noise component, and 3) the fundamental frequency (Fant et al., 1985; Lalwani and Childers, 1991a). The modeling processes for these three types of features are illustrated in the following subsections.

4.1.1 Low Frequency Waveform

The typical waveform and spectrum of the glottal volume velocity have been introduced in Figure 2–3. It seems that the low frequency component dominates the glottal waveform. Several models were developed to describe this low frequency waveform (Rothenberg et al., 1973; Rothenberg, 1981; Fant et al., 1985; Milenkvoic, 1993). Two popular models are: 1) the polynomial model, and 2) the LF model.

4.1.1.1 Polynomial model

To allow for possible distortion in the data recording process, a robust modeling method is recommended (Milenkovic, 1993; Childers and Hu, 1994). Polynomial fitting is one such modeling procedure. Childers and Hu (1994) adopted a sixth order polynomial to model the differentiated glottal volume-velocity. The waveform of this polynomial model can be written as

$$p(t) = C_0 + C_1\tau + C_2\tau^2 + C_3\tau^3 + C_4\tau^4 + C_5\tau^5 + C_6\tau^6 \qquad 0 < t \leq T \qquad (4\text{–}1)$$

where t is the time variable, $\tau = t / T$, and T is the pitch period. The coefficients, C_i, determine the shape of the differentiated glottal waveform, and they are used as the parameters for the polynomial model.

To derive the polynomial coefficients is an optimization problem, which is usually known as a polynomial fitting algorithm. The fitting algorithm itself is not the main concern of this research; we are interested in the order of polynomial, so that we can simulate the glottal waveform with tolerable distortion.

In our polynomial fitting process, one pitch period of normalized (with unit power) differentiated glottal waveform is used as the target. A polynomial is used to fit the target waveform, and the order of the polynomial is in the range from two to nine. To reduce abrupt changes in the glottal waveforms between adjacent pitch periods the polynomial coefficients are lowpass filtered by the filter H(Z).

$$H(Z) = \frac{0.75\ Z}{Z - 0.25} \tag{4–2}$$

To determine the proper order for polynomial fitting, a distortion measure, DT, is defined as,

$$DT = \sqrt{\frac{1}{M}\sum_{i=1}^{M} | \text{dgvv}(i) - p(i) |^2} \tag{4–3}$$

where M is the total number of samples in one pitch period, dgvv(i) is the ith sample of the original differentiated glottal waveform, and p(i) is the ith sample of the modeled waveform. Based on such a definition, a normalized "error vs polynomial order" plot is shown in Figure 4–1. An abrupt distortion decrement is observed when the polynomial order is changed from 3 to 4. Another obvious decrement occurred when the polynomial order is increased from 5 to 6. Figure 4–2 illustrates an example of the target waveform and its associated model waveforms. As we can see that the modeled polynomial waveforms follow the target waveform well when the polynomial order is no less than four.

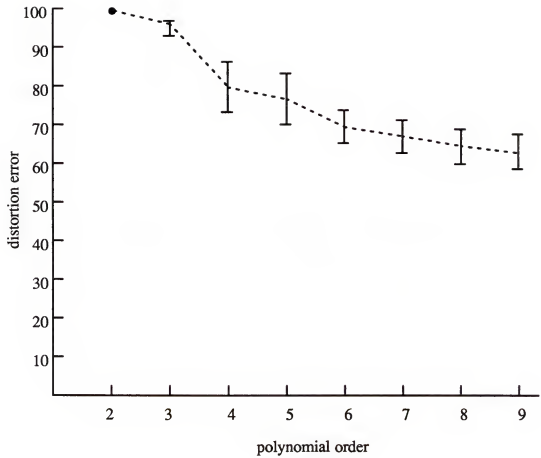


Figure 4–1. The averaged "error vs. polynomial order" plot.
The maximum error is normalized to 100.
The vertical bar denotes the standard deviation.

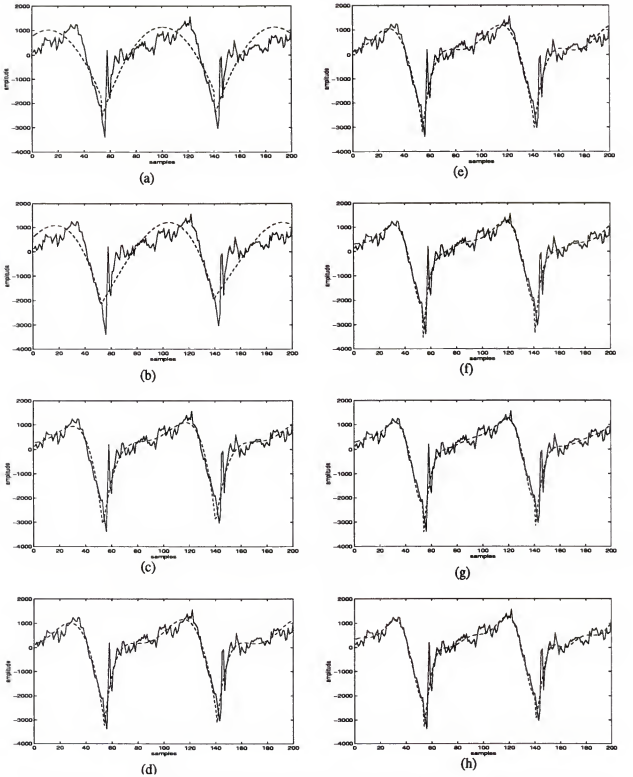


Figure 4–2. The differentiated glottal waveform and its polynomial model waveforms. The solid line is the original waveform and the dashed line is the modeled waveform. (a) order = 2; (b) order = 3; (c) order =4; (d) order =5; (e) order = 6; (f) order = 7; (g) order = 8; (h) order = 9.

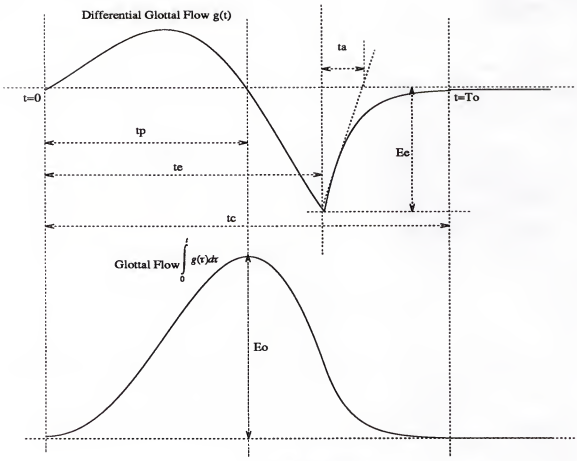
4.1.1.2 LF model

Although the polynomial glottal source model is robust, the lack of a physical correlation of the model parameters (polynomial coefficients) is troublesome. There are several acoustic models whose parameters are closely related to the acoustic features of the glottal volume–velocity, and hence are more applicable than the polynomial model for psychoacoustic studies. The LF model is one such model.

The waveform of the LF model and its integral are shown in Figure 2–4 (Fant et al., 1985; Fant and Lin, 1987). For convenience this figure is reproduced in Figure 4–3. The LF model consists of two segments. The first segment is an exponentially growing sinusoid, and the second segment is an exponential decaying function. These two segments are separated by the timing parameter t_e, which is also the instant of the maximum glottal closing rate. The parameter t_p denotes the position of peak glottal waveform, t_a is the time constant of the exponential recovery as well as an indication of the abruptness of glottal closure, and t_c marks the instant of glottal closure.

Procedures have been proposed to fit the estimated glottal waveform from the glottal inverse filtering (GIF) with the LF model timing parameters (Gobl, 1988; Childers and Ahn, 1994). A major problem is marking the glottal opening instant. Previous fitting procedures were designed either by interactively guessing the glottal opening instant or by using the EGG signal as auxiliary information (Childers and Krishnamurthy, 1985).

In this research, instead of making use of the EGG signal or interactively selecting the glottal opening instant, a code–word search procedure is developed in order to find the best set of LF parameters for each pitch period of the differentiated glottal waveform. A forty entry codebook, shown in Table 4–1, provides a possible set of LF timing parameters (t_p, t_e, t_a, and t_c). The design of the code–book is presented in Appendix. Before the search process, an optional preprocess approximates the original differentiated glottal waveform with a sixth order polynomial is provided. Such preprocessing is actually a lowpass



Glottal Parameters

$$g(t) = \begin{cases} E_e e^{\alpha t} \sin \omega_g t & 0 < t \leq t_e \\ -\frac{E_e}{\varepsilon t_a}\left[e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)}\right] & t_e \leq t \leq t_c \leq T \end{cases}$$

with the following restrictions

$$\int_0^T g(t) = 0 \qquad \omega_g = \frac{\pi}{t_p}$$

$$\varepsilon t_a = 1 - e^{-\varepsilon(t_c - t_e)} \qquad E_o = -\frac{E_e}{e^{\alpha t_e} \sin \omega_g t_e}$$

T = *pitch period*

Figure 4–3. The LF model. The model waveform and its integral waveform.

Table 4–1. The code–book for the LF model timing parameters.

Codeword	t_p (%)	t_e (%)	t_a (%)	t_c (%)
c1	54.401	87.277	6.510	100.00
c2	56.286	84.052	6.263	100.00
c3	62.399	95.496	2.529	100.00
c4	52.628	80.602	5.809	99.645
c5	90.984	93.443	2.377	100.00
c6	77.420	84.781	0.980	89.759
c7	49.742	57.340	4.557	78.377
c8	68.526	75.308	2.551	87.433
c9	73.864	92.122	0.585	94.999
c10	68.627	94.222	1.934	99.886
c11	60.705	76.310	2.609	88.776
c12	51.785	61.020	2.494	73.151
c13	50.437	57.896	0.090	59.463
c14	62.866	69.208	0.161	71.087
c15	23.333	33.333	9.667	80.000
c16	27.721	38.161	3.947	57.173
c17	33.072	43.804	3.120	58.812
c18	33.700	42.422	4.158	62.434
c19	37.253	48.189	3.535	65.234
c20	33.168	48.486	4.039	67.973
c21	36.810	50.215	3.923	69.155
c22	41.748	55.945	3.066	70.736
c23	35.377	50.354	3.003	64.884
c24	37.533	61.311	3.192	76.772
c25	52.626	78.065	2.998	92.104
c26	45.357	70.366	4.345	90.645
c27	37.844	57.730	4.686	80.063
c28	46.464	66.287	4.179	86.125
c29	45.599	63.277	2.868	77.184
c30	40.215	59.280	3.505	76.231
c31	59.371	92.425	2.170	99.623
c32	40.945	60.649	0.806	65.192
c33	40.204	55.734	0.596	59.172
c34	42.478	60.572	1.839	69.680
c35	42.205	56.954	1.046	62.253
c36	44.765	59.399	1.604	67.271
c37	27.871	39.237	0.497	42.324
c38	32.886	39.548	0.557	42.663
c39	34.256	46.863	1.240	53.097
c40	35.884	47.390	2.053	57.300

filtering of the original glottal waveform. The search process exhaustively compares the target waveform with the modeled waveform to determine the codeword with the smallest "acoustic distance." The fast Fourier transform (FFT) spectrum is used as the acoustic distance measure so that a time alignment of glottal opening instants is not required. A peak power normalization is done prior to the spectral comparison. Since the characteristics of the glottal waveform do not change rapidly, the search process is executed once every four frames (pitch periods), so that the same set of LF parameters, once chosen, are assigned for next four consecutive pitch periods.

Three segments of typical glottal waveforms are shown in Figure 4–4 along with the respective best fit coded waveforms. These waveforms are considered to be representative of three voice types, i.e. modal, vocal fry, and breathy (Childers and Ahn, 1994). The spectral distance, SD, is defined as

$$SD = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |df(i) - xf(i)|^2} \qquad (4\text{–}4)$$

where N is the total number of spectral components, df(i) is the normalized spectral component of the modeled waveform, and xf(i) is the normalized spectral component of the target waveform. The SDs corresponding to the waveforms in Figure 4–4 are summarized in Table 4–2. The best fit SD is at least five times less than the average SD and ten times less than the worst fit SD.

Table 4–2. Summary of the spectral distance for the glottal waveform in Figure 4–4.

	Best fit SD	Average SD	Worst fit SD
Glottal waveform 1	19.2	98.5	193.2
Glottal waveform 2	12.3	88.0	232.3
Glottal waveform 3	33.6	192.6	629.5

Figure 4–5 illustrates an example that illustrates the effect of the preprocessing. When the preprocessing is included in modeling, the modeled waveform has a larger t_a value than its counterpart. This result is reasonable, because the larger t_a is, the more the high frequency components are reduced in the glottal waveform.

In addition to the above code–word search procedure, six sets of typical LF parameters are preset for six voice types. These typical parameter values are obtained from previous research (Ahn, 1991; Gobl, 1988).

4.1.2 Modeling of the Noise Component

As shown in the previous subsection, both the polynomial and the LF models can be used to simulate the low frequency component of the glottal waveform. By subtracting the low frequency component from the differentiated glottal waveform, a noise–like signal, as illustrated in Figure 4–6 (a), is observed. This noise–like signal, which is attributed to turbulent noise, is important for the naturalness of synthetic speech (Holmes, 1976; Klatt, 1980; Kasuya et al., 1986; Lalwani and Childers, 1991a). The intensity, spectrum–shaping, and timing information are the general features for the turbulent noise (Klatt and Klatt, 1990; Childers and Lee, 1991; Lalwani and Childers, 1991a).

The design for generating turbulent noise is shown in Figure 4–6 (b). This procedure makes use of a segment of white gaussian noise along with six parameters. The parameters are defined as follows:

1. SNR: The power ratio between the low frequency component and the aspiration noise.
2. amp1: The amplitude modulation index 1, and $0 \leqq \text{amp1} \leqq 1.0$.
3. amp2: The amplitude modulation index 2, and $0 \leqq \text{amp2} \leqq 1.0$.
4. offset: The duration for amp1. The offset starts from the glottal opening instant.

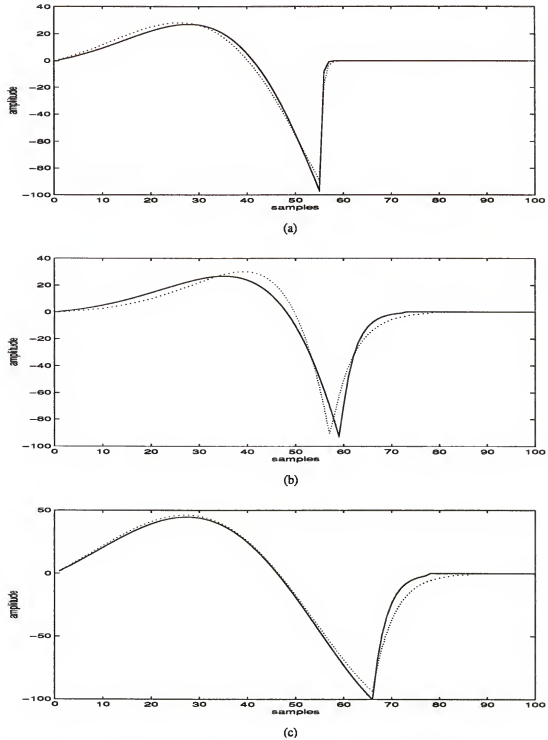


Figure 4–4. The original excitation waveform and its best fit coded waveform. The solid line indicates the original waveform, and the dotted line denotes the best fit coded waveform.
(a) t_p=41.3, t_e=55.4, t_a=0.4, t_c=58.2 (modal);
(b) t_p=48.1, t_e=59.6, t_a=2.7, t_c=72.0 (vocal fry);
(c) t_p=46.2, t_e=66.0, t_a=2.7, t_c=77.1 (breathy).

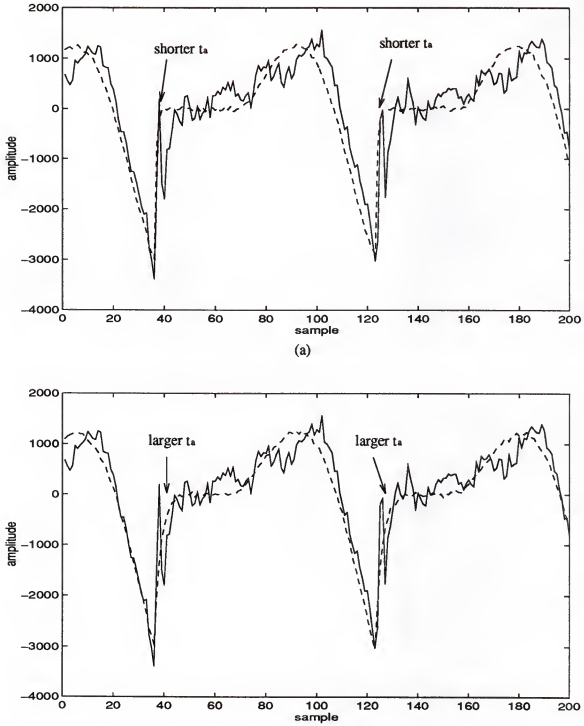
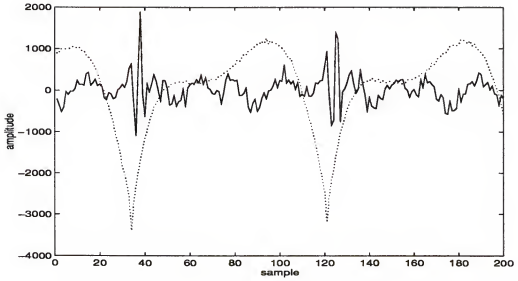
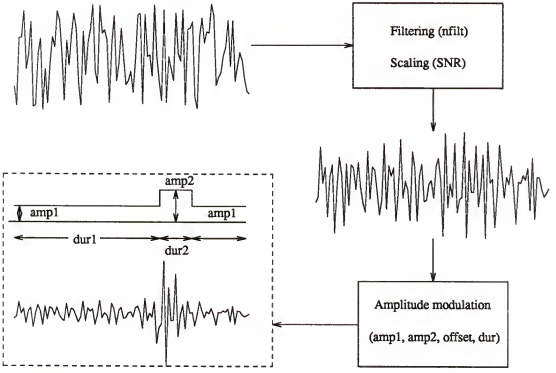


Figure 4–5. The differentiated glottal waveform. The solid line is the original waveform and the dashed line is the modeled waveform.
(a) The modeled waveform is obtained without preprocessing;
(b) The modeled waveform is obtained with preprocessing.



(a)



(b)

Figure 4–6. The turbulent noise. (a) A segment of turbulent noise (solid line) derived by subtracting the approximated LF model waveform (dotted line) from the differentiated glottal waveform; (b) Illustration of the model for generating turbulent noise.

5. dur: The duration for amp2. The dur starts from the end of the offset, and the offset plus the dur should not be greater than the pitch period. If the offset plus the dur is greater than the pitch period, the dur will automatically be reduced to the pitch period minus the offset. If the offset plus the dur is less than the pitch period, the rest of the duration will adopt amp1 as modulation index.
6. nfilt: The spectral tilt of the aspiration noise, $-1.0 \leqq \text{nfilt} \leqq 1.0$.

The white noise with specific length (pitch period) is first filtered according to the parameter nfilt. Actually, nfilt is the coefficient of a first order discrete–domain filter. The filtered noise is then scaled by the parameter SNR. The scaled noise is amplitude modulated by the parameters amp1, amp2, offset, and dur, which simulates the amplitude fluctuations of the turbulent noise due to the variations in airflow and glottal area during vocal fold vibration (Childers and Lee, 1991).

The typical values for the above parameters are speaker dependent and vocal quality dependent. Generally, for modal voices, the SNR value is small, typically about 0.25%. The turbulent noise is high–pass filtered, and the largest magnitude occurs around the vicinity of the glottal closure instant (Childers and Lee, 1991).

4.1.3 Variations of the Fundamental Frequency Contour

In addition to the low frequency component and the turbulent noise of the glottal source, the fundamental frequency (f0) is an important acoustic factor for assessing individual differences in voice quality perception (Kreiman et al., 1992; Kreiman et al., 1994). Therefore, a model that is able to vary the fundamental frequency contour is needed by a synthesis system to produce various types of voices.

The present model for varying the fundamental frequency contour is shown in Figure 4–7. The original fundamental frequency contour can be obtained by analyzing a segment of speech signal. This f0 contour can be scaled up or down by multiplying by a constant

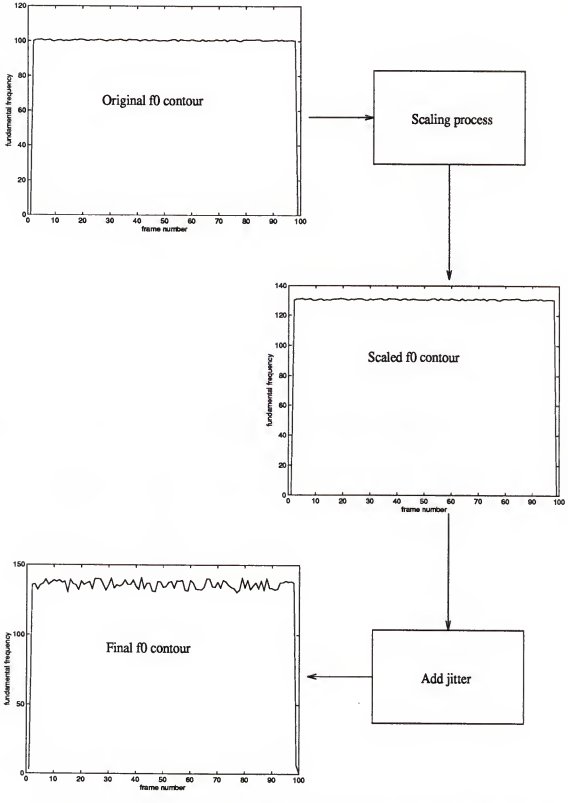


Figure 4–7. Illustration of the model for varying the fundamental frequency contour.

factor, "f0_scale." The jitter component is introduced in order to fluctuate the scaled f0 contour. The jitter range is set proportional to the averaged fundamental frequency.

The modeling processes for the above three types of features can be manipulated by a graphic user interface (GUI) software, `src_model`. The usage of this GUI is presented in Chapter 5.

4.2 Vocal Quality and Glottal Model Parameters

Laver and Hanson (1981) have classified vocal quality into six major types: modal, vocal fry, breathy, whisper, falsetto, and harsh. These six types of vocal quality are sometimes referred to as six voice types. Qualitative explanation of the phonated features of various voice types has been discussed in the area of linguistics (Ladefoged, 1971; Laver, 1980; Laver and Hanson, 1981; Kreiman et al., 1994), and summarized as follows:

1. Modal: The acoustic characteristics for modal voice are: a long opening phase and a rapid closing phase, medium pitch period, low turbulent noise, low pitch perturbation, and low amplitude perturbation, medium source spectral tilt but pitch dependent (Childers and Lee, 1991).
2. Vocal fry: Vocal fry is defined perceptually to be a low-pitched (usually 30 – 90 Hz) , rough sounding phonation. The period to period variation of the fundamental frequency (jitter) is quite high. The glottal spectrum of vocal fry voice falls off less steeply than that of other voices (Monsen and Engebretson, 1977). The glottal waveform has sharp, short pulses followed by a long closed glottal interval. The glottal opening phase may have one, two, or three opening/closing phases (Hollien and Michel, 1968).
3. Breathy: Breathy is slightly audible friction. The degree of breathy severity is inversely proportional to the length of the closed glottal phase (Eskenazi et al., 1990). High pitch perturbation is an acoustic feature for breathy voice (Eskenazi and Childers, 1990).

4. Falsetto: Falsetto is perceived as a flute-like tone that is sometimes breathy. The glottis with gradual opening and closing phases often remains slightly apart. The glottal pulse has a short or non-closed phase. Hollien and Michel (1968) stated that the fundamental frequency is high (275–634 for males).
5. Whisper: The acoustic spectrum of whisper is similar to that of breathy voice. Both breathy voice and whisper involve the presence of audible friction. In other words, the transition from breathiness to whisper is an auditory continuum. Whether the vocal folds vibrate or not during whispering is still in dispute (Zemlin,1981; Kaplan, 1971).
6. Harsh : Harshness involves large variations in the vibratory pattern of the vocal folds. The acoustic characteristics of harsh voice are mainly described as an irregularity of glottal waveform and with spectral noise.

In the last few years more concern has been given to the quantitative relationships between the glottal features and specific vocal quality. The formant synthesis is one of the methods that has been used for the above application (Lalwani, 1991; Childers and Wu, 1990). In this section we will first review the research concerning the glottal characteristics of the different voice types. Then we will study a voice conversion process in order to extend our knowledge about synthesizing all six voice types by using the formant–based LP synthesizer.

4.2.1 Previous Research

Based on the hypothesis that the vocal quality is primarily affected by the shape of the glottal pulse (Rosenberg, 1971), Childers and Lee (1991) have adopted the analysis, synthesis and perception strategy to examine the source–related features for four voice types: modal, vocal fry, breathy, and falsetto. In their experiment, both the speech and EGG signals are used in the analysis phase. Four factors were found to be important for

characterizing the glottal excitations for the four voice types: the glottal pulse width, the glottal pulse skewness, the abruptness of glottal closure, and the turbulent noise component. Based on the perceptual assessment of the synthesized speech, the source–related features and their typical values for the four types of voices are summarized.

By a systematic variation of the time domain factors of a proposed glottal source model, Lalwani (1991) has used a formant synthesizer to validate certain hypotheses concerning the relationships between the time domain glottal factors and various voice types. Both aspiration noise and fundamental frequency perturbation are included as critical glottal factors. Through a listener's evaluation of the vocal quality of the sustained vowel /i/, the preferred values of the glottal factors for modal, vocal fry, and breathy voice are summarized in Table 4–3.

Childers and Ahn (1994) have used the glottal inverse filtering and the glottal model fitting processes to model features of the glottal volume–velocity waveform for three voice types: modal, vocal fry, and breathy voice. The statistical analysis (ANOVA) shows that there was a difference in three of the four LF model parameters for the three voice types. The most significant LF model parameters for each voice type were also determined by a linear regression analysis between the four LF model parameters and a formal rating by a listening test of the quality of the three voice types. Table 4–4 illustrates the mean values and standard deviations for the LF model parameters for modal, vocal fry, and breathy voices.

4.2.2 Voice Conversion

For certain voice types, such as whisper and harshness, it is difficult to detect the synchronous timing information (pitch period and glottal closure instant) as well as other acoustic features via a typical speech analysis procedure (Lalwani and Childers, 1991b). Therefore, the glottal inverse filtering and the glottal model fitting processes might not be

Table 4–3. Preferred values of the glottal factors for three voice types
All parameters are defined in Section 4.1.
(From Lalwani, 1991)

Parameter / Quality	f0 (Hz)	t_p%	t_e%	t_a%	t_c%	amp1	amp2	offset%	dur%	jitter%
Modal	100	45.0	60.0	0.5	N/A	0.0	1.0	50	50	5
Vocal fry	40	15.0	20.0	0.1	N/A	0.0	1.0	50	50	15
Breathy	100	60.0	90.0	8.0	N/A	0.0	1.0	50	50	5

Table 4–4. Mean Values and standard deviations for the LF model parameters for the modal, vocal fry, and breathy voices (From Childers and Ahn, 1994).

Parameter / Quality	f0 (Hz)	t_p%	t_e%	t_a%	t_c%	SQ_{LF} [2]
Modal	118.63 (11.16)[1]	41.34 (5.49)	55.50 (7.77)	0.41 (0.92)	58.17 (8.84)	2.80 (1.33)
Vocal fry	101.26 (30.82)	48.08 (17.81)	59.55 (17.76)	2.69 (2.20)	72.00 (21.66)	2.34 (1.08)
Breathy	114.28 (27.96)	46.21 (11.01)	66.04 (16.14)	2.70 (2.08)	77.12 (15.27)	1.62 (0.71)

[1] The value in parenthesis is the standard deviation
[2] $SQ_{LF} = t_p / (t_c - t_p)$

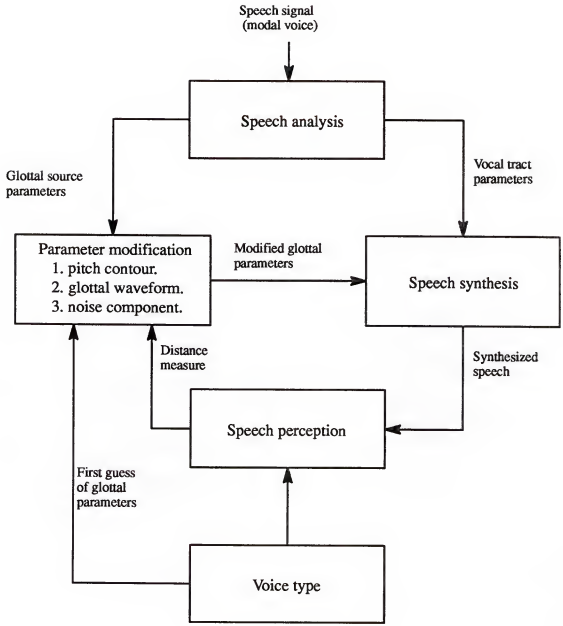


Figure 4–8. The voice type conversion procedure.

Table 4–5. The preferred values of the glottal parameters for six voice types.

Quality / Glottal factor	Modal	Vocal fry	Breathy	Whisper	Falsetto	Harshness
f0_scale	1.0	0.6	1.0	0.9	3.0 [1]	1.0
jitter %	2.0	10.0	5.0	2.0	2.0	10.0
t_p %	45.0	20.0	50.0	50.0	50.0	25.0 [2]
t_e %	60.0	25.0	80.0	80.0	80.0	30.0 [2]
t_a %	0.5	0.2	8.0	8.0	8.0	1.0 [2]
t_c %	65.0	35.0	100.0	100.0	100.0	50.0 [2]
SNR (dB)	40.0	20.0	20.0	–20.0	50.0	10.0
amp1 %	0.0	0.0	100.0	100.0	0.0	100.0
amp2 %	100.0	100.0	100.0	100.0	0.0	100.0
offset %	50.0	20.0	50.0	50.0	50.0	50.0
dur %	50.0	20.0	50.0	50.0	50.0	50.0

[1] To maintain the duration of the converted falsetto voice equal to the duration of the original modal voice, the f0_scale factor is achieved by repeating one pitch period of the glottal waveform three times, and then decimating the resultant waveform by three.
[2] The LF parameters for the harsh voice vary according to the formula, y=(1+d)x, where x is the typical parameter value, d is a random number in the range from –0.2 to 0.2, and y is the final parameter value for a specific frame.

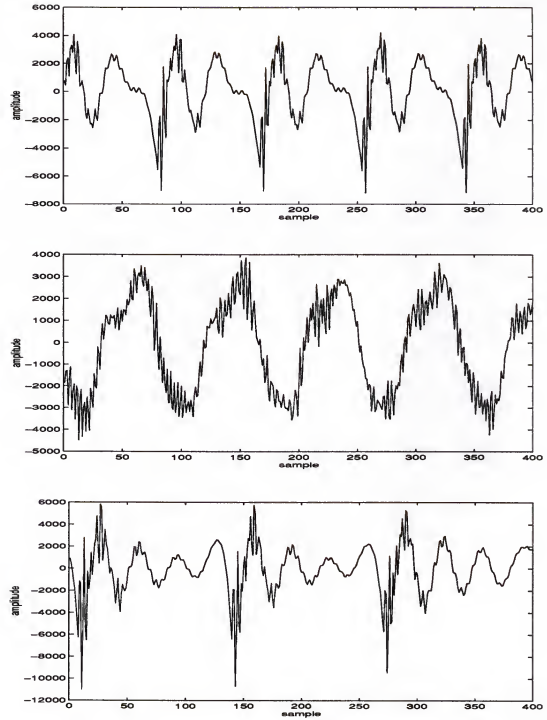


Figure 4–9. Synthesized waveforms of sustained vowel /i/.
(a) The synthesized waveform of modal quality;
(b) The synthesized waveform of breathy quality;
(c) The synthesized waveform of vocal fry quality;
(d) The synthesized waveform of falsetto quality;
(e) The synthesized waveform of whisper quality;
(f) The synthesized waveform of harsh quality.

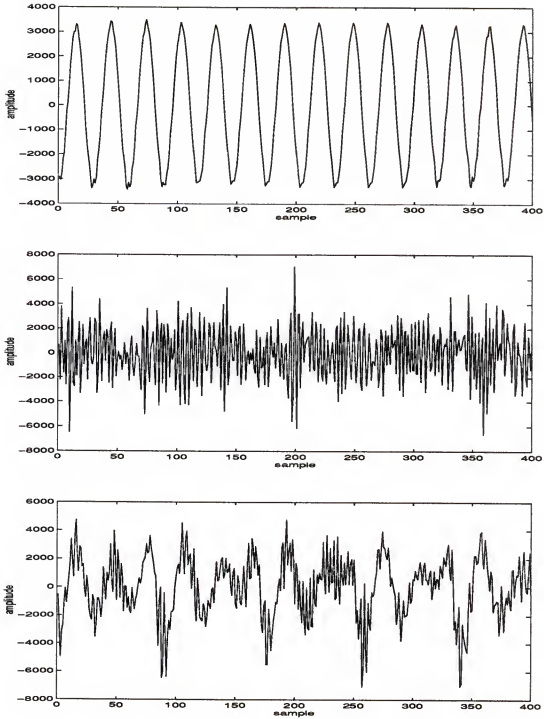


Figure 4–9. Continued.

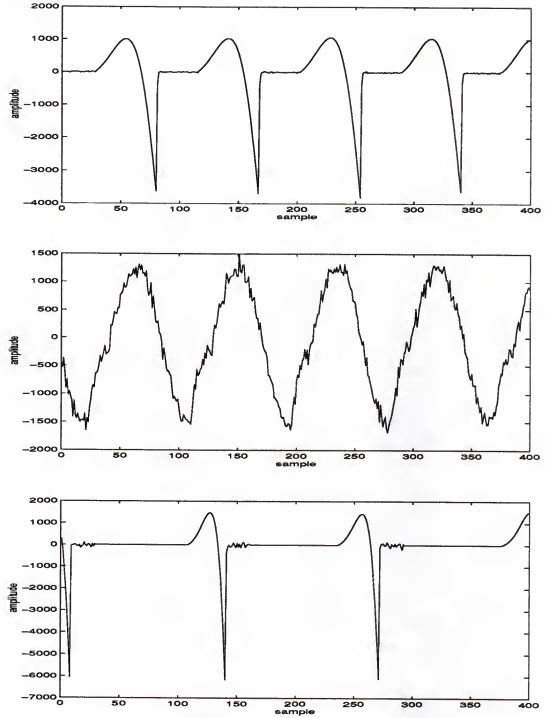


Figure 4–10. Excitation waveforms for the sustained vowel / i /.
(a) The excitation waveform for modal voice;
(b) The excitation waveform for breathy voice;
(c) The excitation waveform for vocal fry voice;
(d) The excitation waveform for falsetto voice;
(e) The excitation waveform for whisper voice;
(f) The excitation waveform for harsh voice.

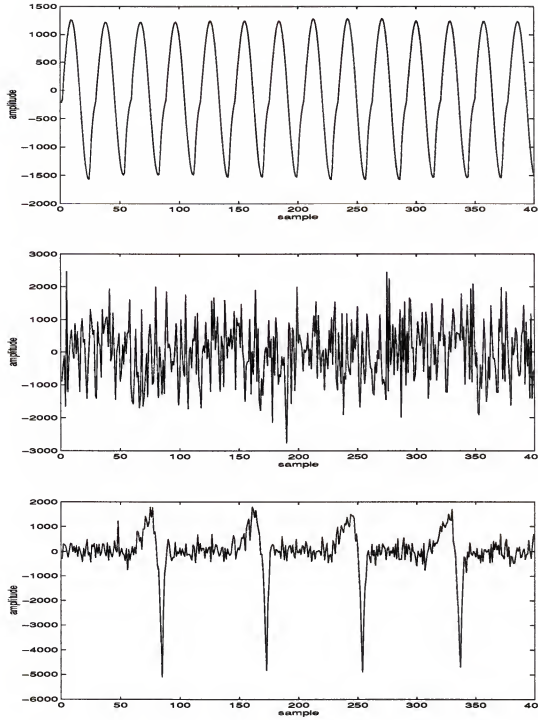


Figure 4–10. Continued.

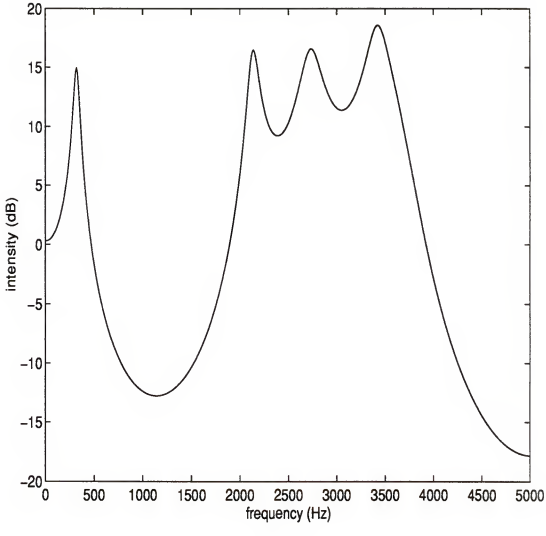


Figure 4–11. The vocal tract's frequency response for the sustained vowel / i /.

adequate to extract the correct parameter values. The voice conversion method makes use of a speech analysis/synthesis system to convert one voice type to another by using an appropriate glottal excitation pulse model. This has been used to explore the essential glottal features for various voice types (Childers et al, 1989b).

The procedure for voice conversion is illustrated in Figure 4–8. In this procedure, a segment of speech of "modal" quality is analyzed. The analyzed vocal tract parameters are directly sent to the formant–based LP synthesizer, whereas the glottal source parameters, such as pitch contour, low frequency glottal waveform, and noise component are modified according to the voice type being designated. The speech perception block provides a distance measure that determines when the converted synthetic speech is satisfactory (Kreiman et al., 1992; Kuwabara et al., 1991). Therefore, a synthesis/perception loop is formed to find the appropriate values for the glottal parameters in this voice conversion process. In current research, the synthesis/perception loop is executed in a trial and error manner. The synthesized speech is perceptually assessed by the user, who determines when the conversion process is completed.

By applying several sustained vowels (phonated by male speakers) through the voice conversion process, Table 4–5 summarizes the preferred values of certain glottal factors for six voice types. The definitions of these glottal factors can be found in Section 4.1. Note that, the conversion rules in Table 4–5 are valid only when the original speech is a male modal voice.

According to the above rules, Figure 4–9 demonstrates one segment of synthesized sustained vowel / i / for each type of vocal quality. The corresponding excitation waveforms are illustrated in Figure 4–10, and the frequency response of the vocal tract transfer function is shown in Figure 4–11. Also based on the conversion rules, an all–voiced sentence "We were away a year ago." of modal quality can be converted to five other voice types. An informal listener evaluation confirmed that the synthesized voices were representative of their natural counterparts.

4.3 Summary

The modeling process for the glottal source was presented in this chapter. The low frequency waveform, the aspiration noise, and the variation of the fundamental frequency (f0) track were separately modeled. The modeling process does not need the EGG signal. Also, the modeling process can directly provide the glottal model parameters to the formant–based LP synthesizer. Based on the glottal modeling results and previous studies concerning the relationships between vocal quality and glottal features, a voice conversion procedure was established, which can be used to verify and expand our knowledge about the essential glottal features for various voice types, such as whisper, falsetto, and harsh voices. While the synthesis/perception loop is preliminary, several types of voices can be generated by the formant–based LP synthesizer. This result implies us that the linkage between the glottal source modeling process and the synthesis system may serve as a basis for further study of: 1) the estimation of parameters for excitation source models for a broad range of voice types, and 2) a data base of different voice types to be used in training a speech recognition system (Childers and Ahn, 1994).

CHAPTER 5

GRAPHIC USER INTERFACE

The purpose of this chapter is to introduce two graphic user interfaces (GUIs) for the formant–based LP synthesizer and the glottal source modeling process. These interfaces can assist the user to 1) assign values to the synthesis and glottal modeling parameters, 2) load and store data from files of specific formats, and 3) execute the synthesis and glottal modeling processes without memorizing sophisticated commands.

5.1 Speech Synthesis

A typical synthesis procedure is shown in Figure 5–1. This procedure is accomplished by two phases: 1) specifying the parameters, and 2) processing the parameters. The GUI described below stresses the first phase of synthesis.

For the formant–based LP synthesizer, two kinds of specification files have to be built: 1) the general specification file, and 2) the source specification file. All the vocal tract parameters and control parameters are specified in the general specification file. The source specification file contains the excitation source parameters. The following subsections introduce the way of generating these two files through the use of GUIs. The "Main function window," as shown in Figure 5–2, is the window that provides main functions in the formant–based LP synthesizer.

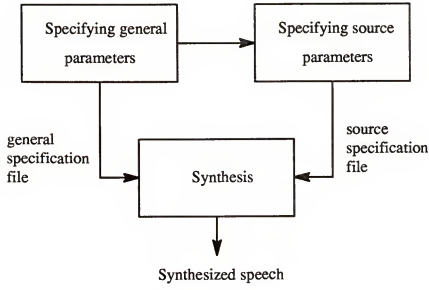


Figure 5–1. A typical synthesis procedure.

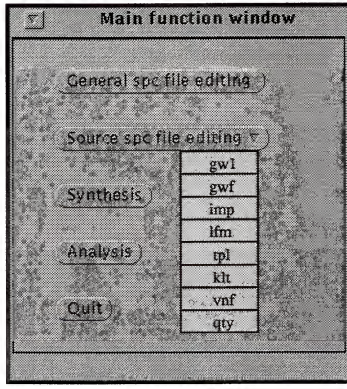
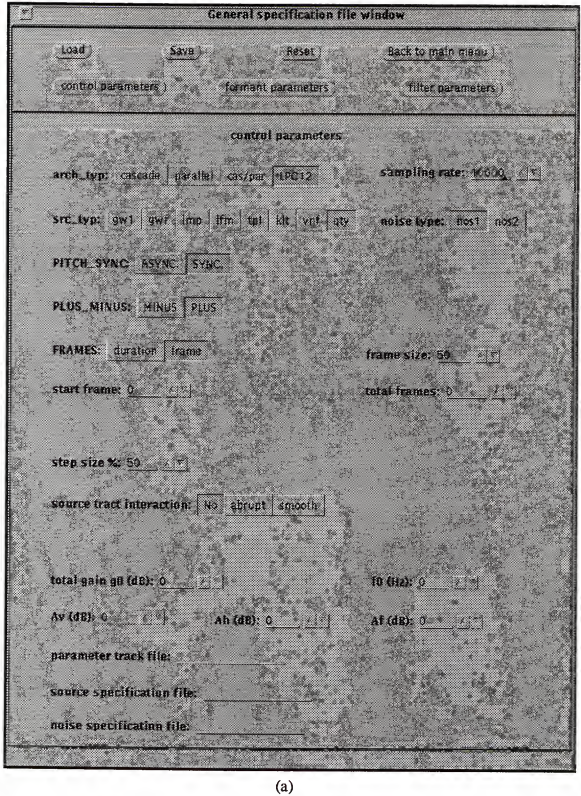
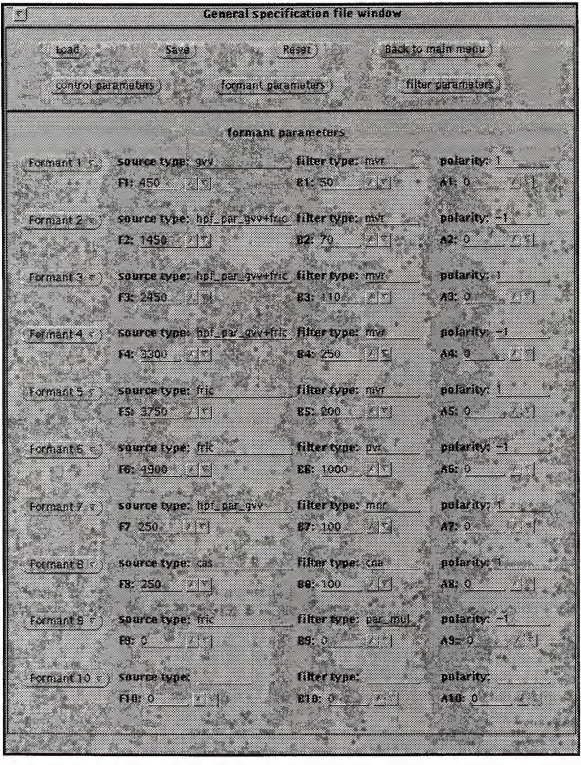


Figure 5–2. The window for selecting main function in the formant–based LP synthesizer.



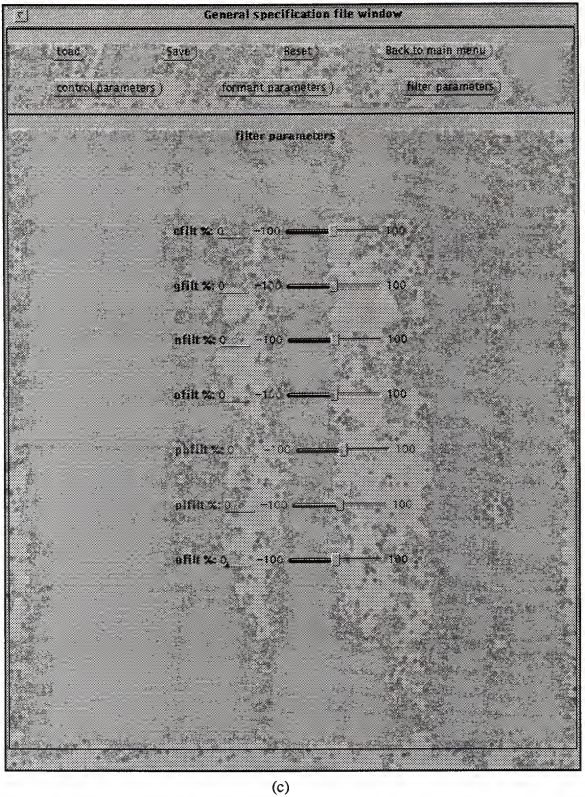
(a)

Figure 5–3. The windows for specifying the general specification file. (a) The control parameters; (b) The formant parameters; (c) The filter parameters.



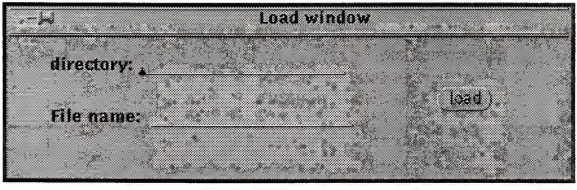
(b)

Figure 5–3. Continued.

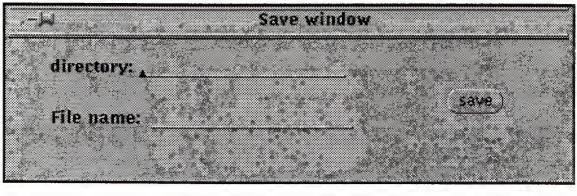


(c)

Figure 5–3. Continued.



(a)



(b)

Figure 5–4. Windows for the file input / output.
(a) Pop–up window for loading a file;
(b) Pop–up window for saving a file.

5.1.1 General Parameters

The "General spc file editing" button in the main window opens the "General specification file window" shown in Figure 5–3. The function buttons are shown in the top frame. The "Load" button opens a pop–up "Load window" shown in Figure 5–4. Users can load any general specification file by filling in the directory and filename. Once the file is loaded, the values of synthesis parameters are assigned accordingly. The "Save" button opens another pop–up window for storing a general specification file. The "Reset" button invokes the program that sets all synthesis parameters to their default value. The default values are stored in a file, called `def_new.d`. The "Back to main menu" button closes this "General specification file window."

There are three windows that constitute the GUI for specifying the general parameters. Since only one window can be monitored at a time, the "control parameters," "formant parameters," and "filter parameters" buttons determine the window being monitored. The synthesis parameters are shown in the lower panel of the "General specification file window." Because the development of the formant–based LP synthesizer is founded on Lalwani's (1991) flexible formant synthesizer, the parameters shown in the interface window might be employed by either previous or current or both synthesizers. We would like to stress the parameters that are utilized by the formant–based LP synthesizer.

1. The control parameters (Figure 5–3 (a)):

The mutual exclusive push–button "arch_typ" provides four choices for determining the synthesis scheme and the vocal tract architecture. The first three architectures are Lalwani's implementation, and the fourth one belongs to the formant–based LP synthesizer. The numerical field "sampling rate" determines the sampling rate in Hz of our digital synthesis environment. The mutual exclusive push–button "src_typ" provides eight

types of excitation sources. The buttons "gwf," and "qty" invoke the two most often used source models for the formant–based LF synthesizer. The GUIs for specifying these two glottal sources will be illustrated later. The push–button "PITCH_SYNC" determines the way that the synthesis parameters are renewed. They can be updated either pitch synchronously ("SYNC.") or by frame basis ("ASYNC."). The push–button "FRAMES" denotes the time unit being used for synthesis. It can be either in msec ("duration") or in frame ("frame"). The numerical fields, "start frame" and "total frames," illustrate the first synthesis frame and the total number of synthesis frames, respectively. The fields, "total gain g0 (dB)," "f0 (Hz)," "Av (dB)," and "Af (dB)," specify the parameters defined in Section 2.3. Notice that the parameters specified through the general specification file are constants. For the situation that needs the time–varying synthesis parameters, a parameter track file that contains the time–varying parameters is employed. The filename of this parameter track file is specified in the "parameter track file" field. The "source specification file" field denotes the file that contains the glottal source parameters.

2. The formant parameters (Figure 5–3 (b)):

The numerical fields, "F1," "F2," "F3," "F4," "F5," "F6," determine the formant frequencies in Hz. The fields, "B1," "B2," "B3," "B4," "B5," "B6," specify the formant bandwidths in Hz. The rest of parameters in this window are not used by our synthesis system.

3. The filter parameters (Figure 5–3(c)):

Seven sliding bars, "cfilt," "gfilt," "nfilt," "ofilt," "phfilt," "plfilt," "ufilt," all ranged from –1 to +1, are used to determine the coefficients of the first order filters introduced in Section 2.3.3.

5.1.2 Glottal Source Parameters

The "Source spc file editing" menu in the main window provides eight kinds of glottal sources. The interfaces for the two most often used glottal sources, "gwf," and "qty," are introduced as follows:

1. The GWF window

The menu item "gwf" in the main window opens the "GWF source window" shown in Figure 5–5. This window allows the user to specify a file that contains a segment of sampled signals. The sampled signals are directly applied to the synthesizer as input excitation. The length of the synthesized speech should be equal to the length of the input excitation.

The "Load," "Save," "Reset," and "Back to main menu" buttons shown in the top panel of the "GWF source window" all work in the same way as they were introduced in the previous subsection. By using these buttons, the user can load or store a source specification file, reset the source parameters that belong to the "gwf" source to their defaults, or close the "GWF source window."

The manipulations of the source parameters are shown in the lower part of the "GWF source window." The numerical field "Scaling Factor %" specifies the percent of scaling factor for the input excitation. The mutual exclusive push–button "F0 effect" determines whether the changes of the fundamental frequency would affect the amplitude of the input waveforms or not. The other mutual exclusive push–button "Type of gain" denotes the strategy being used to interpret the relationship between the input excitation and the gain parameters such as, "av" or "af." For example, the first button "PWR" indicates that for each synthesis frame, the power of the input excitation must be adjusted in order to satisfy the gain parameters. Detailed definition for the gain type was summarized in Lalwani's work (1991). For the "gwf" source, we usually do not intend to modify the input excitation

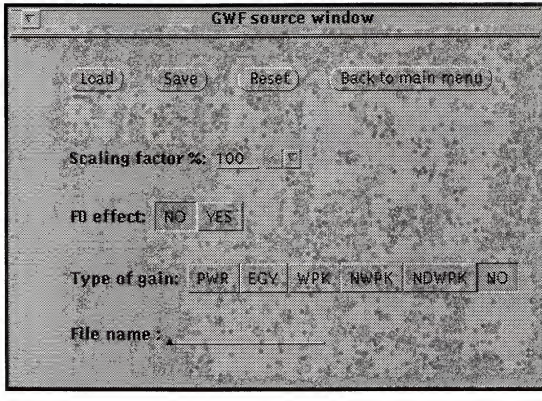
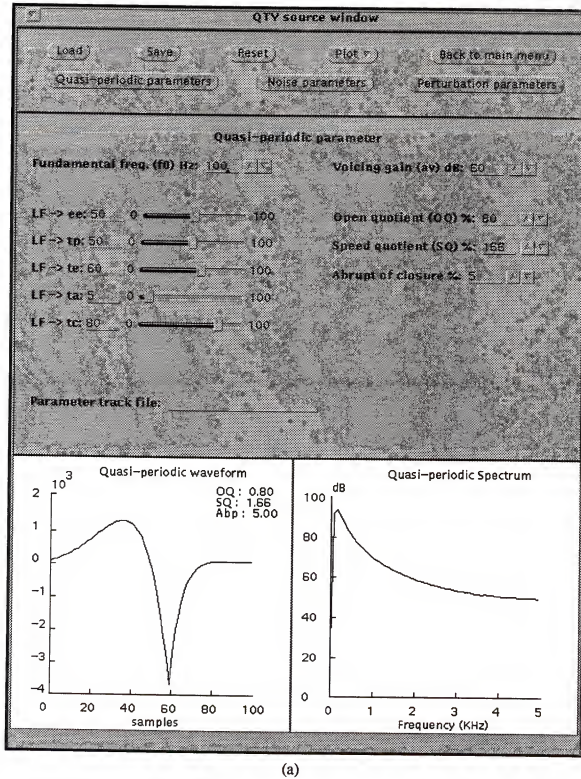
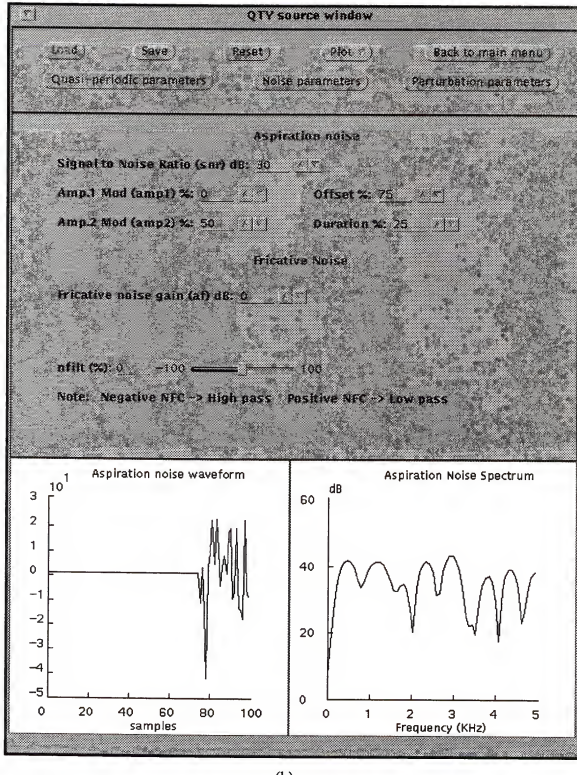


Figure 5–5. The window for specifying the "gwf" source.



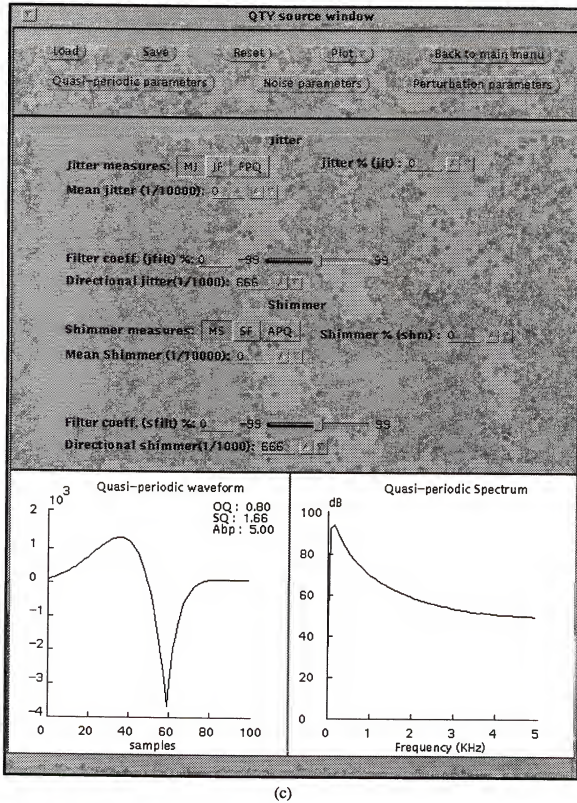
(a)

Figure 5–6. Windows for specifying the "qty" source.
(a) The window for specifying the quasi–periodic parameters;
(b) The window for specifying the noise parameters;
(c) The window for specifying the perturbation parameters.



(b)

Figure 5–6. Continue.



(c)

Figure 5–6. Continue.

obtained from a speech analysis procedure. Therefore, the "NO" button is the default gain type. In other words, the input excitation will not be scaled before the first pass of the synthesis (see Section 2.4.1). The text field "File name" specifies the file that contains the glottal waveforms.

2. The QTY window

The menu item "qty" in the main window opens the "QTY source window" shown in Figure 5–6. These windows allow the user to specify the waveforms and noise components of an extended LF source model. The function buttons, "Load," "Save," are used to handle the I/O of the source specification file. The "Reset" button resets the parameters for the "qty" source model. The "Back to main menu" button closes the "QTY source window." The push–button "Plot" provides the capability to illustrate the modeled glottal waveform and spectrum. The waveform and spectrum are displayed in the bottom panel of the QTY window.

There are three windows that constitute the GUI for the extended LF source model. Since only one window is displayed at a time, the "quasi–periodic parameters," "noise parameters," and "perturbation parameters" buttons determine the window being displayed on the screen.

a. The window for quasi–periodic parameters (Figure 5–6 (a)): The numerical fields, "Fundamental freq. (f0) Hz," and "Voicing gain (av) dB," specify the fundamental frequency and the voiced gain of the voiced sounds, respectively. The sliding bars, "LF–>ee," "LF–>t_p," "LF–>t_e," "LF–>t_a," and "LF–>t_c," are the LF model parameters that determine the shape of the glottal waveform. These LF model parameters can also be specified indirectly by the numerical fields, "Open quotient (OQ) %," "Speed quotient (SQ) %," and "abrupt of closure %." The text field "Parameter track file" denotes the file that contains the time–varying parameter track.

b. The window for noise parameters (Figure 5–6 (b)): The noise parameters include the aspiration noise and fricative noise. The "Signal to Noise Ration (snr) dB" field shows the power ratio between the quasi–periodic components and aspiration noise. The lower the signal to noise ratio is, the higher the aspiration noise will be. The resultant aspiration noise is then amplitude modulated through the use of the numerical fields "Amp.1 Mod (amp1) %," "Amp.2 Mod (amp2) %," "Offset %," and "Duration %." The first two fields specify the amplitude modulation indexes and the last two fields indicate the associated period for the respective modulation index. The numerical field "Fricative noise gain (af) dB" illustrates the power of the unvoiced sounds. The "nfilt (%)" sliding bar, which ranges from –1 to 1, allows users to tilt the spectrum for both the aspiration and fricative noise.

c. The window for perturbation parameters (Figure 5–6(c)): In generating natural sounding sustained vowels, the fundamental frequency (f0) and the power for voiced sound (Av) are usually perturbed frame–by–frame, while the perturbation might be small. The jitter and shimmer are the parameters that determine the maximum range of perturbation with respect to the mean f0 and Av. Since the jitter measures have been defined in several ways (Lalwani, 1991), the mutual exclusive push–button "jitter measures" is used to select the preferred measure for jitter. In Figure 5–6(c), owing to the "JF" button being selected, the "Jitter factor (%)" field is shown in the "perturbation parameters" window instead of other fields ("Mean jitter," or "Freq. perturbation quotient"). The amount of the frequency perturbation can be specified either by the "Jitter % (jit)" field or by the "Jitter factor (%)" field. This is because these two fields are mathematically dependent on each other. The frequency–domain feature of the perturbation is determined by the sliding bar "Filter coeff. (jfilt) %," which ranges from –1 to +1. The positive value of this parameter enforces a slow change of the perturbation sequence, otherwise, rapid perturbation would be expected. The numerical field "Directional Jitter (1/1000)" which characterize the frequency–domain feature of the

differentiated perturbation sequence is closely related to the parameter "jfilt." An empirical formula listed below is used to map the filter coefficient "jfilt" to the directional jitter measure.

$$DJ = 0.0067a^6 - 0.0434a^5 + 0.0037a^4 + 0.0885a^3 - 0.0608a^2 - 0.1623a + 0.6662 \quad (5\text{–}1)$$

where DJ is the directional jitter, and "a" is the filter coefficient.

The process for specifying the shimmer parameters is similar to the one for jitter.

5.1.3 Examples

Two synthesis examples are illustrated below. The first example is to generate a sustained vowel / i / with length of 100 frames, and the second one is to re–synthesize a short word / cheep /. Both examples are presented in the way that first gives the specifications, then introduces the synthesis process, and finally discusses the result.

Example One

a. Specifications

   Except for the glottal source parameters, the other specifications listed below are obtained by analyzing a segment of a sustained vowel / i / spoken by a male speaker. For convenience, this original sustained vowel was called a "target token" in later text.

   i. Total number of synthesis frames is 100.

   ii. Voiced gain "av" is 60 dB.

   iii. Fundamental frequency "f0" is 108 Hz.

   iv. Jitter "jit" is 1%.

   v. The first five formants shown in Table 5–1 are constant throughout the synthesis.

vi. The glottal source parameters are specified according to the typical values suggested by Ahn (1991), and summarized in Table 5–2.

Table 5–1. The first five formants for synthesizing the sustained vowel / i /.

	1st formant	2nd formant	3rd formant	4th formant	5th formant
frequency	316	2115	2704	3286	3550
bandwidth	56	50	500	450	300

Table 5–2. Specified values for the modified LF model parameters.

Parameter	t_p(%)	t_e(%)	t_a(%)	t_c(%)	snr(dB)	nfilt
Value	44	58	1	60	30	–0.99

b. Synthesis process

i. Specify the general parameters in a general specification file

- Move to the working directory and start the GUI by typing "my_fmtsyn."
- Open the "General specification file window."
- Select the "control parameters" button and specify the parameters, "total frame," "Av(dB)," and "f0 (Hz)."
- Insert the "src1.d" in the "source specification file" field.
- Select the "formant parameters" button and specify the first five formants.
- Open the pop–up "Save" window and insert the filename, "spc1.d," in the filename field.
- Close the "General specification file window" by pressing the "Back to main menu" button.

ii. Specify the source parameters in a source specification file

- Open the "QTY source window" by selecting the "qty" button under the "Source spc file editing" menu.
- Select the "Quasi–periodic parameters" button and specify the parameters, "f0," "av," "t_p," "t_e," "t_a," and "t_c."
- Select the "Noise parameters" button and specify the parameters, "snr," and "nfilt."
- Select the "Perturbation parameters" button and specify the parameter "jit."
- Open the pop–up "Save" window and type "src1.d" in the filename field.
- Close the "QTY source window" by pressing the "Back to main menu" button.

iii. Synthesis

- Open the "Synthesis window" shown in Figure 5–7 by selecting the "Synthesis" button in the "main function window."
- Write the "spc1.d" in the "General specification file" field.
- Write the "out1.e" in the "Output speech file" field.
- Execute the synthesis by pressing the "Synthesis" button.
- Close the "Synthesis window" by pressing the "Back to main menu" button.

c. Results

A segment of synthesized speech waveform and its spectrum along with the target token are shown in Figure 5–8. Both the time and frequency domain's features for the synthesized and the target token are similar, except there are more high frequency components in the target token. This phenomenon is probably caused by the insufficient modeling of the high frequency component of the glottal source. The synthesized token sounds natural, and no click or pop noises are observed.

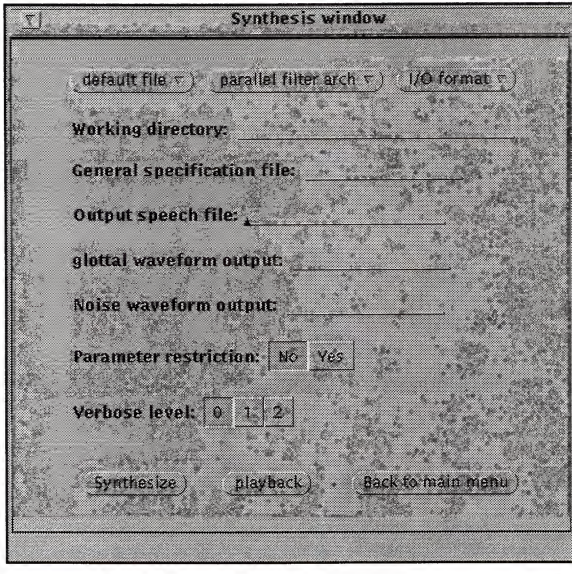


Figure 5–7. The synthesis and playback window.

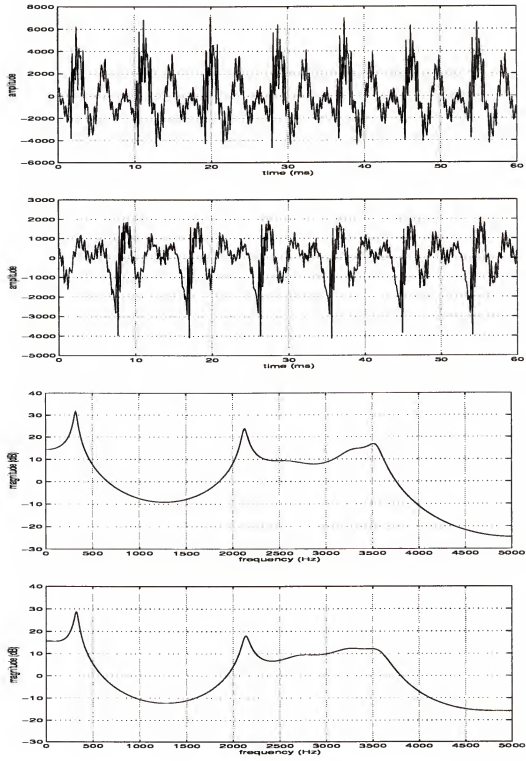


Figure 5–8. Waveforms and spectrum for the sustained vowel / i /.
(a) Target waveforms; (b) Synthesized waveforms;
(c) Target spectrum; (d) Synthesized spectrum.

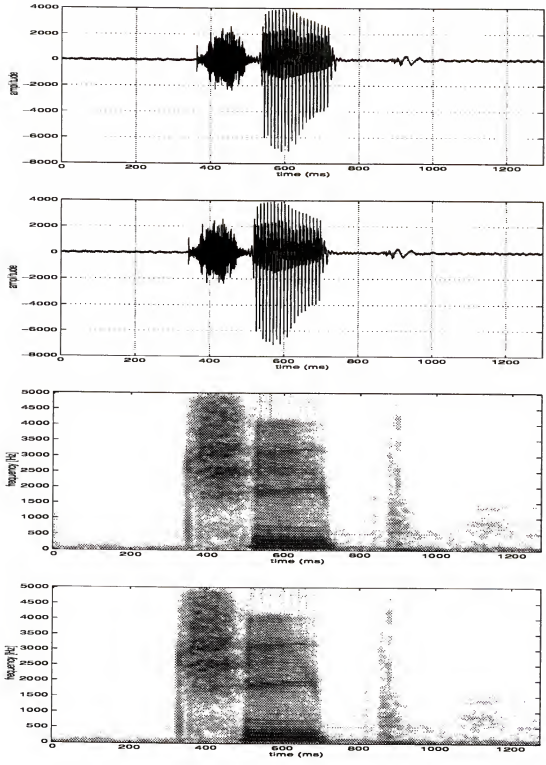


Figure 5–9. Waveforms and spectrograms for the word / cheep /.
(a) Target waveforms; (b) Synthesized waveforms;
(c) Target spectrogram; (d) Synthesized spectrogram.

Example Two

a. Specifications

Another target token "cheep" is analyzed pitch synchronously. The time–varying parameters, such as "av," "af," "f0," the first five formants (for voiced regions), and the LP coefficients (for unvoiced regions), are stored in the file named "exp2_some.d." The estimated glottal waveform from the GIF process (for the target token) is stored in another file named "exp2_src.d."

b. Synthesis process

i. Specify the general parameters in a general specification file

- Move to the working directory and start the GUI by typing "my_fmtsyn."
- Open the "General specification file window."
- Select the "control parameters" button.
- Choose the "gwf" as the source type (in the line of "src_typ").
- Insert the "exp2_some.d" in the "parameter track file" field.
- Insert the "src2.d" in the "source specification file" field.
- Open the pop–up "Save" window and insert "spc2.d" in the filename field.
- Close the "General specification file window"by pressing the "Back to main menu" button.

ii. Specify the source parameters in a source specification file

- Open the "GWF source window" by selecting "gwf" button under the "Source spc file editing" menu.
- Insert the "exp2_src.d" in the "File name" field.
- Open the pop–up "Save" window and insert "src2.d" in the filename field.
- Close the "GWF source window" by pressing the "Back to main menu" button.

iii. Synthesis

- Open the "Synthesis window" shown in Figure 5–7 by pressing the "Synthesis" button in the "main function window."
- Write the "spc2.d" in the "General specification file" field.
- Write the "out2.e" in the "Output speech file" field.
- Execute the synthesis by pressing the "Synthesis" button.
- Close the "Synthesis window" by pressing the "Back to main menu" button.

c. Results

Figure 5–9 illustrates the waveforms and spectrograms of the target and synthetic speech tokens. It is hard to distinguish them by comparing either the envelopes of the waveforms or the frequency distributions of the spectrograms. Notice that the reason for this almost perfect re–synthesis is due to the estimated glottal waveform from the GIF process is used as the excitation. Degradations would be expected if the estimated waveform is modeled.

5.2 Modeling of the Excitation Source

The GUI for the modeling process of the excitation source are presented as follows.

1. The main window:

Figure 5–10 shows the main window for the glottal source modeling process. In this window, the input/output files for the modeling are specified. The input files contain the original speech signal and the waveform from the glottal inverse filtering (GIF) process. Note that the length of these two signals are equal. The field, "Modeled waveform file," specifies the file that contains the modeled glottal waveform. Since through the modeling process the synthesis parameters might change, the "Synthesis parameter file" field denotes

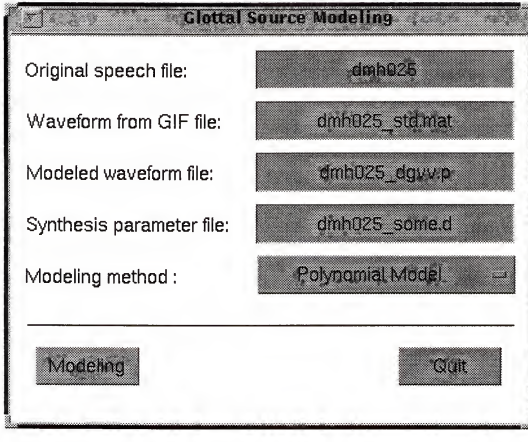


Figure 5–10. Main window for glottal source modeling.

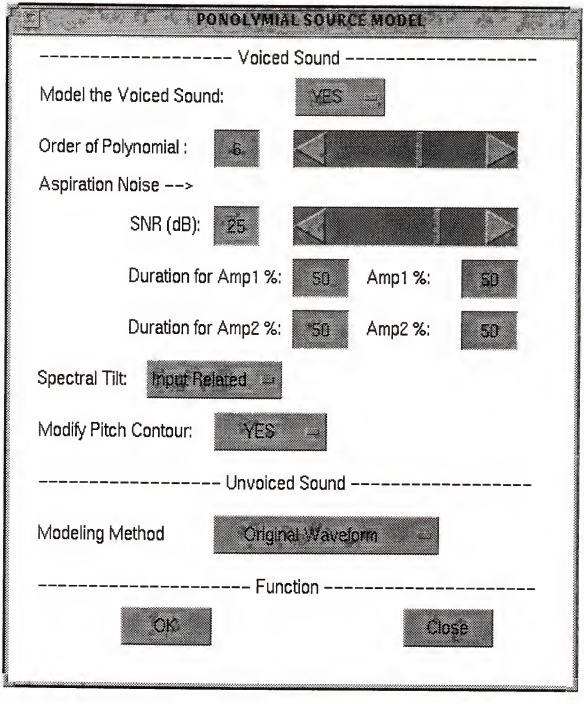


Figure 5–11. The window for the polynomial source model.

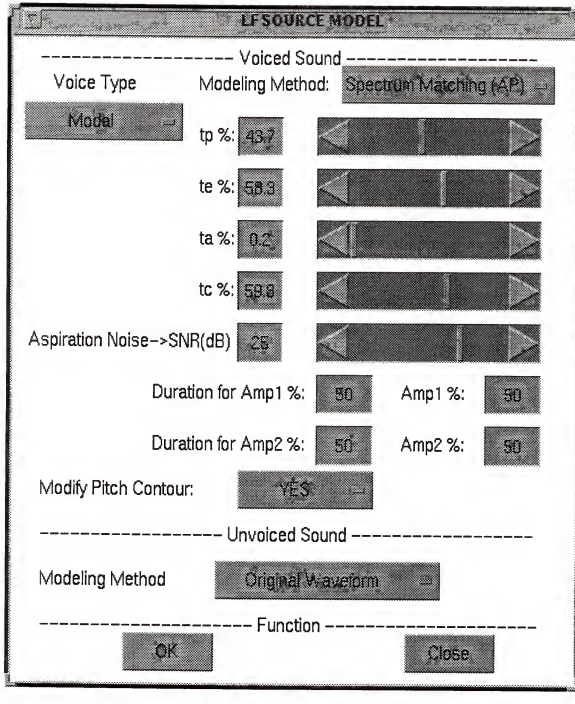


Figure 5–12. The window for the LF source model.

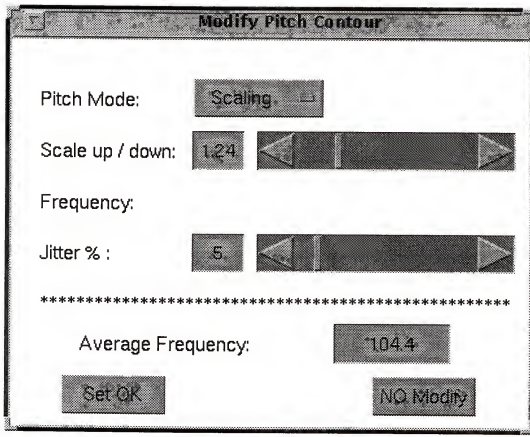


Figure 5–13. The window for varying the pitch contour.

the file that contains the updated synthesis parameters. The selection button, "Modeling method," specifies the model (polynomial or LF model) that simulates the low frequency component of the glottal source. According to the selected modeling method, the "Modeling" button opens the specific modeling window. The "Quit" button closes all the windows for the glottal source modeling.

2. The polynomial source model:

The window for the polynomial source model is shown in Figure 5–11. This window is used not only for simulating the low frequency component of the glottal source but also for specifying the aspiration (turbulent) noise parameters as well as representing unvoiced sounds. The selection button, "Model the Voiced Sound," provides an option for the user to determine whether the voiced sound is to be modeled or not. The sliding bar, "Order of Polynomial," specifies the polynomial order. For the aspiration noise, the sliding bar, "SNR (dB)," controls the power ratio between the low frequency component and the aspiration noise (before amplitude modulation) of the glottal source. The fields, "Duration for Amp1 %," "Amp1," "Duration for Amp2 %," and "Amp2" specify the parameters that amplitude modulates the aspiration noise. With the existence of the spectral difference between the original glottal waveform and the modeled waveform, the selection button, "Spectral Tilt," provides the flexibility to adjust the spectral tilt. The adjustment can either be input related (changed frame–by–frame) or be fixed. The "Modify Pitch Contour" button decides whether the f0 contour will be modified or not. If it is set to "YES," the f0 modification window (Figure 5–11) will be opened. The "Modeling Method" button for the unvoiced sound shows the way of representing the unvoiced sound. When all other parameters are set, the function button, "OK," invokes the modeling process. The other function button, "Close," closes the window for the polynomial source model.

3. The LF source model:

Figure 5–12 illustrates the window for the LF model. Based on the "Modeling Method" button, four methods have been developed to provide an appropriate set of LF parameters for each voiced frame. The methods include "Default," "User Define," "Spectrum Matching," and "Spectrum Matching (LP)." In the "Default" method, one of six sets of typical values for the LF parameters is selected for all modeling frames. The "Voice Type" menu provides the six typical sets of LF parameters. The "User Define" method enhances the flexibility of specifying the LF parameters, and this method is accomplished by manually controlling the sliding bars "t_p %," "t_e %," "t_a %," and "t_c %." The "Spectrum Matching" and "Spectrum Matching (LP)" methods invoke the search process for the best fit LF parameters. The difference between these two methods is that the "Spectrum Matching (LP)" method follows the process of lowpass filtering the glottal waveform and the "Spectrum Matching" method doesn't. The rest of the control buttons or fields in this window work in a similar manner as described for the polynomial source model.

4. Variations of the fundamental frequency contour:

The window for varying the f0 contour is shown in Figure 5–13. The "Pitch Varying Mode" button denotes the manner for modifying the original f0 contour. It can be either "Scaling" or "Constant." If the "Scaling" button is selected, the f0 contour will be scaled depending on the specified value from the sliding bar "Scale up/down." Otherwise, a constant f0 contour, which is set by the use of the sliding bar "Frequency," will replace the original f0 contour. The sliding bar, "Jitter %," specifies the range of the fluctuations that will be added to the scaled f0 contour. The "Average Fundamental Frequency" field is read–only; it shows the average f0 value of the original f0 contour. The function button, "Set OK," invokes the program of varying the f0 contour. The other function button, "Reset," resets the f0 contour to its original value.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

The primary goal of this research was to develop a speech synthesis/analysis system, which is capable of generating high–fidelity synthetic speech as well as a tool for psychoacoustic studies. In order to accomplish this primary goal, a formant–based linear prediction (LP) synthesizer along with a robust speech analysis procedure were developed. One major feature of this synthesis/analysis system is its ability to adapt the formant and linear prediction schemes to represent the voiced and unvoiced sounds, respectively (Olive, 1992). Based on these synthesis and analysis procedures, as well as a knowledge of the relationship between vocal quality and glottal features (Childers and Lee, 1991; Childers and Ahn, 1994), a voice conversion procedure was developed to convert a segment of speech signal of modal quality to five other voice types (vocal fry, breathy, falsetto, whisper, and harsh). The following sections summarize the results of this study and discuss the direction of future work.

6.1 Summary of Results

6.1.1 Speech Synthesizer

The formant–based linear prediction (LP) synthesizer is a realization of the source–tract speech production model (Fant, 1960). Two types of excitation sources: 1)

the voiced source and 2) the unvoiced source, form the source part of this synthesizer (Holmes, 1973; Klatt, 1980; Lalwani and Childers, 1991a). The major feature of this synthesizer is its two way representation for the vocal tract. The advantages of employing two kinds of schemes in one synthesis system are: 1) the formant scheme is physically meaningful for simulating the human speech production system, and 2) the LP scheme is able to reproduce the spectrum of all speech sounds. The vocal tract transfer function was modeled by a twelfth order polynomial (as the denominator of the transfer function) and a constant (as the numerator). Therefore, the vocal tract was represented as an all–pole filter, and this filter was constructed in a direct–I structure, i.e. in a linear prediction format. The twelfth order polynomial for unvoiced sound was obtained by a twelfth order LP analysis (autocorrelation method) using a frame size of 50 samples (5 ms). The polynomial for voiced sounds was generated by multiplying six second order polynomials together, where each polynomial was obtained from its corresponding formant frequency and bandwidth.

Since different schemes (formant and LP representations) were used to model the vocal tract, the spectral difference at the boundaries between the voiced and unvoiced sounds were unavoidable (Olive, 1992). The spectral difference would cause a waveform discontinuity problem. This problem was manipulated by a two–pass synthesis strategy, which controlled the gain of the excitation source and the initial state of the vocal tract filter.

Instead of having several filters in parallel (Verhelst and Nilens, 1986; Lalwani, 1991), the formant–based LP synthesizer uses only one linear prediction filter, whose initial state is controlled, thereby, dealing with the residual energy from the previous synthesis frame. No coefficient sensitivity problem was observed in our synthesis system. Also, the vocal tract filter was stable for voiced sounds, and was determined to be empirically stable for unvoiced sounds when the LP coefficients were rounded to four decimal figures.

A graphic user interface program called `my_fmt.syn` was implemented by the use of `deviguide`, `XView`, and `C` functions. The user can conveniently manipulate the synthesis parameters through this interface software.

The synthesis system can re–synthesize speech almost perfectly when the estimated glottal waveform from the glottal inverse filtering process is used as the excitation. When the modeled glottal waveform is used as the excitation source, the synthesized speech is natural and intelligible.

6.1.2 Analysis Procedure

The proper control of the synthesis parameters is the key factor in producing high quality speech. A two–phase LP–based software, `ana_inface`, that analyzes a segment of speech signal was developed to estimate the synthesis parameters such as the V/U classification, fundamental frequency contour, signal power contour, formants (for voiced sounds), and LP coefficients (for unvoiced sounds), as well as the estimated glottal waveform from glottal inverse filtering to the formant–based LP synthesizer.

In the first analysis phase, a two–way classification algorithm was developed to segment speech signals into voiced and unvoiced regions. The first reflection coefficient and the energy of the prediction error were the criteria for classification. In general, a large first reflection coefficient (above 0.2) and a large prediction error energy (above 10^7) denoted a voiced region. Since it is unusual to have only one voiced frame surrounded by unvoiced frames and vice versa, a correction process was invoked to manipulate this kind of situation. Experimental results showed that the two–way classification was acceptable compared to other research (Childers et al., 1989a; Lee, 1992).

Also in the first analysis phase, a pitch detection procedure was developed by the use of the prediction error waveform. This pitch detection procedure is a two–pass process. The first pass determines the average pitch period (fundamental frequency) and the second

pass locates the glottal closure instants (GCIs). Less than 2% average drift of detecting the fundamental frequency has been determined. An interactive software called `modgci` was implemented to correct the possible error in locating GCIs. This software can zoom in and zoom out the prediction error waveform as well as the GCI sequence, and provide functions to add or remove one or several GCIs.

The objective of the second phase LP analysis is to extract the formants. Three methods: 1) quadratic interpolation of the asynchronous LP coefficient, 2) conventional covariance LP analysis, and 3) closed–phase covariance LP analysis were implemented to achieve this objective. By inspecting the differences between true and estimated formants and the similarity between the true and estimated excitation waveforms (for synthetic speech), the closed–phase covariance method provides the exact formants and the exact excitation waveform when the excitation waveform has a longer closed–phase duration (above 30% of the pitch period). When the closed–phase duration went shorter, the estimated first formant might be shifted but within acceptable range (70 Hz).

6.1.3 Voice Conversion

Assuming that the variation of vocal quality is mainly affected by the features of the glottal source, a voice conversion process that reproduces the vocal tract component, but varies the glottal features was proposed to generate sounds of various voice types. Three categories of glottal features: 1) the low frequency waveform, 2) the turbulent noise, and 3) the variation of the fundamental frequency contour, were inspected in the conversion process.

According to the voice conversion experiments we determined: 1) the characteristics of the fundamental frequency contour are critical for producing voice types such as vocal fry, and falsetto, 2) harsh voiced is characterized by a large fluctuation of the fundamental frequency (jitter), as well as the abrupt time–varying feature of the glottal source

parameters (e.g. the LF parameters), 3) the intensity of the turbulent noise plays a dominating role for whisper voice, 4) breathy voice has features that include a wide glottal pulse with a strong turbulent noise, 5) vocal fry voice has a narrow glottal pulse.

Generally, the conversion procedure provided a systematic method for examining the relationships between vocal quality and glottal features. It also provided the capability to establish a data base for different voice types, which can be used in training a speech recognition system (Childers and Ahn, 1994).

6.2 Future Work

1. Model for the unvoiced source

Although we adopted the same order of polynomial to describe the vocal tract for both voiced and unvoiced sounds, unvoiced sounds did not sound as good as voiced sounds when the modeled excitation sources were employed. In the current system, more than ten parameters were used to model the voiced excitation source and only two parameters (intensity and spectral tilt) were employed to model the unvoiced counterpart. Inadequate modeling of the unvoiced excitation might be the cause for degrading the sounding quality (Stevens, 1993b; Stevens et al., 1992). In Childers and Hu's (1994) research, they used the same length of polynomial for the vocal tract, but emphasized the modeling of the unvoiced excitation (encoding the unvoiced excitation in a 256 entries stochastic book). They successfully produced the unvoiced and mixed sounds. This provides us one possible way to enhance the quality for the synthesized unvoiced sound.

2. Formant estimation

In the current analysis procedure, the closed–phase covariance method provides a reliable solution to the formant estimation problem, especially when the closed–phase

duration of the glottal pulse is longer than 30% of the pitch period. When the closed–phase duration is less than 30% of the pitch period, an interactive procedure that permits a manipulation of the formants is suggested (Gobl, 1988). For this interactive procedure, the criterion that determines when the best estimated formants is achieved should be explored in future research.

3. Speech perception and voice conversion

In voice conversion process, an objective criterion that functions as an alternative for speech perception should be studied in order to accelerate the conversion process and narrow the range of parameters for each voice type (Bladon and Lindblom, 1981; Boff et al., 1986; Stephanie, 1986; Kitawaki and Nagabuchi, 1988). Besides, a formal listening test along with a statistic analysis are suggested to verify the validness for the voice conversion rules (Childers and Ahn, 1994).

4. Extension and application

Currently the voice conversion process is applied only to male voices. Different conversion rules are to be expected for female voices. It is possible to characterize gender differences in the glottal waveform for various voice types (Monsen and Engebretson, 1977; Kuwabara and Ohgushi, 1984; Murry and Singh, 1980).

In addition to the glottal source parameters, the vocal tract parameters can be manipulated by our synthesis/analysis system as well. Since the features of the glottal source and the vocal tract are both involved in speech studies such as gender conversion, speaker identification, and speech recognition, this synthesis/analysis system can serve as a tool for future applications (Childers et al., 1990; Childers et al., 1989b; Karlsson, 1992).

APPENDIX

CODEBOOK DESIGN FOR THE LF PARAMETERS

The purpose of this appendix is to illustrate the concept and procedure for designing a codebook for the LF parameters. Four timing parameters (t_p, t_e, t_a, t_c) constitute the vector space for the codebook(Fant et al., 1985).

An infinite number of glottal waveforms could be generated if the model parameters are varied in a continuous manner. To group the glottal waveforms into a finite number of clusters and represent each cluster by one set of specific values for the LF model parameters is the objective of this appendix. This objective can be achieved by a vector quantization method.

Vector Quantization and Vector Space

Quantization converts a continuous-amplitude signal into one of a set of discrete-amplitude signals. The difference between the original continuous-amplitude signal and its discrete counterpart is called the quantization error. In a multi-dimensional space, the continuous-amplitude signal can be quantized separately (scalar quantization) or jointly (vector quantization). Usually for the same number of quantization levels, the vector quantization (VQ) scheme is better than the scalar quantization (SQ) in reducing quantization error (Linde et al., 1980; Makhoul et al., 1985).

Theoretically, the vector space can be constituted by an infinite number of samples. However, it is reasonable to make use of the features of the glottal waveform to reduce the

amount of samples in the vector space. The vector space of this research is formed by one thousand sets of LF parameters, and each set of parameters is obtained by gathering the analysis results from Ahn's research (1991), who analyzed a large data base of speech and EGG signals for three voice types (modal, vocal fry, and breathy), and extracted the LF parameters for each analysis frame. To be specific, the problem is to quantize the sets of LF parameters into L quantized levels.

Distortion Measure

The quantization error, QE(X,Y), is defined as the difference between a four–dimensional vector sample, $X=[x_1\ x_2 \cdots x_4]^T$, and its closest quantized vector, $Y=[y_1\ y_2 \cdots y_4]^T$. The concept of designing the codebook is to find a set of quantized vectors (of L levels), $\{ Y_1, Y_2, \cdots Y_L \}$, which make the overall quantization error OQE a minimum. Thus,

$$OQE = \sum_{i=1}^{M} QE(X_i , Y_j) \qquad 1 \le j \le L \qquad (A\text{–}1)$$

where M is the total number of samples in the vector space, and Y_j is the closest quantized vector to the vector sample X_i.

A common measure for the quantization error is the mean-square error (MSE), and is defined as,

$$QE_{MSE}(X, Y) = \frac{1}{N}(X - Y)^T(X - Y) \qquad (A\text{–}2)$$

where N is the length of the vector X, and the superscript ^{T} is the transpose operation for vector.

The Mahalanobis measure, QE_{mhb}, is a variation of the MSE measure. It is defined as,

$$QE_{mhb}(X, Y) = (X - Y)^T \Gamma^{-1} (X - Y) \qquad (A-3)$$

where Γ is the covariance matrix for the random vector X.

$$\Gamma = E[(X - \overline{X})(X - \overline{X})^T], \qquad \overline{X} = E[X] \qquad (A-4)$$

and E[ ] is the expectation operator.

Assuming real data, Γ is

$$\Gamma = \begin{bmatrix} \overline{t_p^2} - \overline{t_p}^2 & \overline{t_p\, t_e} - \overline{t_p}\,\overline{t_e} & \overline{t_p\, t_a} - \overline{t_p}\,\overline{t_a} & \overline{t_p\, t_c} - \overline{t_p}\,\overline{t_c} \\ \overline{t_p\, t_c} - \overline{t_p}\,\overline{t_e} & \overline{t_e^2} - \overline{t_e}^2 & \overline{t_c\, t_a} - \overline{t_e}\,\overline{t_a} & \overline{t_e\, t_c} - \overline{t_e}\,\overline{t_c} \\ \overline{t_p\, t_a} - \overline{t_p}\,\overline{t_a} & \overline{t_e\, t_a} - \overline{t_c}\,\overline{t_a} & \overline{t_a^2} - \overline{t_a}^2 & \overline{t_a\, t_c} - \overline{t_a}\,\overline{t_c} \\ \overline{t_p\, t_c} - \overline{t_p}\,\overline{t_c} & \overline{t_e\, t_c} - \overline{t_e}\,\overline{t_c} & \overline{t_a\, t_c} - \overline{t_a}\,\overline{t_c} & \overline{t_c^2} - \overline{t_c}^2 \end{bmatrix} \qquad (A-5)$$

$$\overline{t_p^2} = \frac{1}{N}\sum_{i=1}^{N} t_p(i)^2 \qquad (A-6)$$

$$\overline{t_p\, t_c} = \frac{1}{N}\sum_{i=1}^{N} t_p(i) t_e(i) \qquad (A-7)$$

Since the Mahalanobis measure includes the second order statistics of the data samples, it was adopted as the quantization error measure in this research.

Codebook Design

A non–uniform binary tree, as shown in Figure A–1, was employed as the structure for the codebook. The symbols, V_i, were used to represent the intermediate codewords at the branching points. The other symbols, Y_i, were adopted to denote the final codewords for the codebook. Based on this structure and the Mahalanobis measure, we are going to split the data samples into L groups, where each group is represented by a set of LF parameters of particular values. The splitting process is proceeded from the top of the binary tree to the bottom.

At each branching point, the splitting process is initialized by finding two initial vectors as follows:

1. Calculate the arithmetic mean vector, X_0, by

$$X_0 = \frac{1}{K}\sum_{i=1}^{K} X_i \qquad (A-8)$$

where K is the total number of samples in the current group, and X_i denotes the individual sample.

2. Find the first initial vector, X_{f1} that fulfills the following conditions:

$$EQ_{mhb}(X_o, X_{f1}) \geq EQ_{mhb}(X_o, X_i) \text{ for all } X_i, \text{ where } 1 \leq i \leq K, \text{ and } X_{f1} \in X_i \qquad (A-9)$$

3. Find the second initial vector, X_{f2} that fulfills the following conditions:

$$EQ_{mhb}(X_{f1}, X_{f2}) \geq EQ_{mhb}(X_{f1}, X_i) \text{ for all Xi, where } 1 \leq i \leq K, \text{ and } X_{f2} \in X_i \qquad (A-10)$$

Once the initial vectors are obtained, the following iterative procedure splits the samples and finds the optimal intermediate codewords for the two split groups of samples.

1. Split the samples. If $EQ_{mhb}(X_{f2},X_i) \geq EQ_{mhb}(X_{f1},X_i)$, X_i belongs to group 1, otherwise, X_i belongs to group 2.
2. Calculate the mean vectors, X_{01}, and X_{02}, in the following manner, and use them to replace X_{f1} and X_{f2}, respectively.

$$X_{01} = \frac{1}{K_1} \sum_{i=1}^{K_1} X_{i1} \qquad \text{where } K_1 \text{ is the total number of samples in group 1, and } X_{i1} \text{ is the member in group 1.} \qquad (A-11)$$

$$X_{02} = \frac{1}{K_2} \sum_{i=1}^{K_2} X_{i2} \qquad \text{where } K_2 \text{ is the number total of samples in group 2, and } X_{i2} \text{ is the member in group 2.} \qquad (A-12)$$

3. Calculate the quantization errors, EQ_{1j}, and EQ_{2j}, for the jth iteration.

$$EQ_{1j} = \sum_{i=1}^{K_1} EQ_{mhb}(X_{i1}, X_{f1}) \qquad \text{where } K_1 \text{ is the total number of samples in group 1, and } X_{i1} \text{ is the member in group 1.} \qquad (A-13)$$

$$EQ_{2j} = \sum_{i=1}^{K_2} EQ_{mhb}(X_{i2}, X_{f2}) \qquad \text{where } K_2 \text{ is the total number of samples in group 2, and } X_{i2} \text{ is the member in group 2.} \qquad (A-14)$$

4. Define $EQ_j = EQ_{1j} + EQ_{2j}$, as the total quantization error after the jth iteration. If EQ_j minus EQ_{j-1} is less than 0.1, stop the iteration, and the X_{f1} and X_{f2} from the last iteration are the optimal intermediate codewords. Otherwise, go back to step 1 and do the iteration again.

The rules listed below are adopted to determine whether an intermediate codeword is a final codeword or not. In other words, the rules determine where to terminate the splitting process.

1. If the quantization error for the current branching point is less than 15, the splitting process is terminated and the intermediate codeword is a final codeword.
2. If the quantization error for the current branching point is larger than 15 but less than 30, however, the splitting process can not provide an improvement greater than 5, the splitting process is terminated and the intermediate codeword is a final codeword.

3. For other cases, the splitting process is needed and the intermediate codeword is not a final codeword.

Based on such rules, a 40 entry codebook was developed and shown in Table 4–1. The average quantization error (Mahalanobis distance) is 0.3314 for the training data.

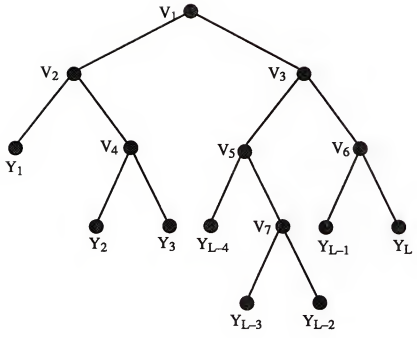


Figure A–1. The non–uniform binary tree structure for codebook design.

REFERENCES

Ahn, C. (1991). A study of voice types and acoustic variabilities: analysis by synthesis, Ph.D. dissertation, University of Florida.

Alku, P. (1992). "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," Speech Commun., 11, 109–118.

Allen, E. L., and Hollien, H. (1973). "A laminagraphic study of pulse (vocal fry) register phonation," Folia Phoniatrica 25, 241–250.

Almeida, L. J., and Silva, F. M. (1984). "Variable–frequency synthesis: An improved harmonic coding scheme," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 27.5.1–27.5.4.

Almeida, L. J., and Tribolet, J. M. (1982). "Harmonic coding: a low bit–rate, good quality speech coding technique," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 1664–1667.

Atal, B. S. and Hanauer, S. L. (1971). "Speech analysis and synthesis by linear prediction of the speech wave," J. Acoust. Soc. Am., 50(2), 637–655.

Atal, B. S., and Rabiner, L. R. (1976). "A pattern recognition approach to voiced–unvoiced–silence classification with applications to speech recognition," IEEE Trans. Acoust., Speech, Signal Process., 24(3), 201–212.

Bergstrom, A., and Hedelin, P. (1989). "Code–book driven glottal pulse analysis," Proc. IEEE Int. Conf., Acoust., Speech, Signal Process., 53–56.

Bladon, R. A. W., and Lindblom, B. (1981). "Modeling the judgement of vowel quality difference," J. Acoust. Soc. Am., 69(5), 1414–1422.

Boff, K. R., Kaufman, L., and Thomas, J. P. (1986). Handbook of perception and human performance, Wiley, New York, 15–1 to 15–20.

Carlson R. (1993). "Models of speech synthesis," Speech Trans. Lab. –Q. Prog. Status Rep. (Royal Institute of Technology, Stockholm, Sweden), 1, 1–13.

Carlson R., Granström, B., and Karlsson, I. (1991). "Experiments with voice modelling in speech synthesis," Speech Commun., 10, 481–489.

Chan, K. (1989). Modeling of vocal fold vibration, Ph.D. dissertation, University of Florida.

Childers, D. G. and Ahn, C. T. (1994). "Modeling the glottal volume–velocity waveform for three voice types," J. Acoust. Soc. Am., in press.

Childers, D. G. and Hu, T. H. (1994). "Speech synthesis by glottal excited linear prediction," J. Acoust. Soc. Am., 96(4), 2026–2036.

Childers, D. G. and Krishnamurthy, A. K. (1985). "A critical review of electroglottography," CRC Critical Reviews in Biomedical Engineering, 12(2), 131–164.

Childers, D. G. and Lee, C. K. (1991). "Vocal quality factors: analysis, synthesis, and perception," J. Acoust. Soc. Am., 90(5), 2394–2410.

Childers, D. G., Hanh, M., and Larar, J. N. (1989a). "Silent and voiced / unvoiced / mixed excitation (four–way) classification of speech," IEEE Trans. Acoust., Speech, Signal Process., 37(11), 1771–1774.

Childers, D. G., and Wu, K. (1990). "Quality of speech produced by analysis–synthesis," Speech Commun., 9, 97–117.

Childers, D. G., Wu, K., and Hicks, D. M. (1987). "Factors in voice quality: acoustic features related to gender," Proc. IEEE Int. Conf., Acoust., Speech, Signal Process., 293–296.

Childers, D. G., Wu, K., Hicks, D. M., and Yegnanarayana, B. (1989). "Voice conversion," Speech Commun., 8, 147–158.

Childers, D. G., Principe, J. C., Ting, Y. T., and Lee, K. (1993). "Adaptive WRLS–VFF for speech analysis," submitted to IEEE Trans. on Speech and Audio.

Deller, J. R., Proakis, J. G., and Hansen, J. H. L. (1993). Discrete time processing of speech signals. Macmillan, New York.

Dudley, H. (1936). "Synthesis speech," Bell Labs. Record, 15, 98–102.

Eggen, J. H. (1992). On the quality of synthetic speech, evaluation and improvement, Ph.D. dissertation, University of Eindhoven, The Netherlands.

Eskenazi, L., Childers, D. G., and Hicks, D. M. (1990). "Acoustic correlates of vocal quality," J. Speech Hear. Res., 33, 298–306.

Fant, G. (1956). "On the predictability of formant levels and spectrum envelopes from formant frequencies," in For Roman Jacobson, Mouton and Co., 's–Gravenhage, The Netherlands, 109–120.

Fant, G. (1960). Acoustic theory of speech production, Mouton and Co., 's–Gravenhage, The Netherlands.

Fant, G. (1979). "Glottal source and excitation analysis," Speech Trans. Lab. –Q. Prog. Status Rep. (Royal Institute of Technology, Stockholm, Sweden), 1, 85–107.

Fant, G. (1980). "Voice source dynamics," Speech Trans. Lab. –Q. Prog. Status Rep. (Royal Institute of Technology, Stockholm, Sweden), 2–3, 17–37.

Fant, G. (1982). "The voice source: acoustic modeling," Speech Trans. Lab. –Q. Prog. Status Rep. (Royal Institute of Technology, Stockholm, Sweden), 4, 28–48.

Fant, G. (1993). "Some problems in voice source analysis," Speech Commun., 13, 7–22.

Fant, G., and Ananthapadmanabha, T. V. (1982). "Truncation and superposition," Speech Trans. Lab. –Q. Prog. Status Rep. (Royal Institute of Technology, Stockholm, Sweden) 2–3, 1–17.

Fant, G., and Lin, Q. G. (1987). "Glottal source – vocal tract acoustic interaction," Speech Trans. Lab. –Q. Prog. Status Rep. (Royal Institute of Technology, Stockholm, Sweden), 1, 13–27.

Fant, G., and Lin, Q. G. (1991). "Comments on glottal flow modelling and analysis," in Vocal fold physiology, acoustic, perceptual, and physiological aspects of voice mechanisms, edited by J. Gauffin and B. Hammarberg, Singular Publishing Group, San Diego, CA, 47–56.

Fant, G., Liljencrants, J., and Lin, Q. G. (1985). "A four parameter model of glottal flow," Speech Trans. Lab. –Q. Prog. Status Rep. (Royal Institute of Technology, Stockholm, Sweden), 4, 1–13.

Flanagan, J. L. (1957). "Note on the design of terminal analog speech synthesizers," J. Acoust. Soc. Am., 29, 306–310.

Flanagan, J. L., Coker, C. H., and Bird, C. M. (1962). "Computer simulation of a formant vocoder synthesizer," J. Acoust. Soc. Am., 35, 2003(A).

Flanagan, J. L., and Ishizaka, K. (1978). "Computer model to characterize the air volume displaced by the vibrating vocal cords," J. Acoust. Soc. Am., 63, 1559–1565.

Flanagan, J. L., Ishizaka, K., and Shipley, K. L. (1975). "Synthesis of speech from a dynamic model of the vocal cords and vocal tract," Bell System Tech. J., 54(3), 485–505.

Flanagan, J. L., Ishizaka, K. L., and Shipley, K. L. (1980). "Signal models for low bit–rate coding of speech," J. Acoust. Soc. Am., 68(3), 780–791.

Fujisaki, H., and Ljungqvist, M. (1986), "Proposal and evaluation of models for the glottal source waveform," Proc. IEEE Int. Conf., Acoust., Speech, Signal Process., 1605–1608.

Fujisaki, H., and Ljungqvist, M. (1987), "Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform," Proc. IEEE Int. Conf., Acoust., Speech, Signal Process., 637–640.

Gavidia–Ceballos, L., and Hansen, J. H. L. (1994). "Direct speech feature estimation using an iterative EM algorithm for vocal cancer detection," submitted to IEEE Trans. Biomed. Engineering.

Gobl, C. (1988). "Voice source dynamics in connected speech," Speech Trans. Lab. –Q. Prog. Status Rep. (Royal Institute of Technology, Stockholm, Sweden), 1, 123–159.

Gobl, C., and Chasaide, A. N. (1992). "Acoustic characteristics of voice quality," Speech Commun., 11, 481–490.

Gold, B., and Rabiner, L. R. (1968). "Analysis of digital and analog formant synthesizers," Trans. Audio Electroacoustics, AU–16, 81–94.

Gopinath, B., and Sondhi, M. M. (1970). "Determination of the shape of human vocal tract from acoustic measurements," Bell System Tech. J., 49, 1195–1214.

Gupta, S. K., and Schroeter, J. (1993). "Pitch–synchronous frame–by–frame and segment–based articulatory analysis by synthesis," J. Acoust. Soc. Am., 94(5), 2517–2530.

Hahn, M. (1989). Silence and voiced–unvoiced–mixed excitation classification of speech with applications: a two–channel and a one–channel, Ph.D. dissertation, University of Florida.

Hamlet, S. L. (1981). "Ultrasound assessment of phonatory function," Conf. on Assessment of Vocal Pathology Reports, 11, 128–140.

Hollien, H., and Michel, J. F. (1968). "Vocal fry as a phonatory register," J. Speech Hear. Res., 11, 600–604.

Holmes, J. N. (1973). "The Influence of the glottal waveform on the naturalness of speech from a parallel formant synthesizer," IEEE Trans. Audio Electroacoustics, AU–21, 298–305.

Holmes, J. N. (1976). "Formant excitation before and after glottal closure," Proc. IEEE Int. Conf., Acoust., Speech, Signal Process., 39–42.

Holmes, J. N. (1983). "Formant synthesizers: cascade or parallel," Speech Commun., 2(4), 251–274.

Holmes, W. J., Holmes, J. N., and Judd, M. W. (1990). "Extension of the bandwidth of the JSRU parallel formant synthesizer for high quality synthesis of male and female speech," Proc. IEEE Int. Conf., Acoust., Speech, Signal Process., 313–316.

Hsieh, Y. F. (1994). A flexible and high quality articulatory speech synthesizer, Ph.D. dissertation, University of Florida.

Hu, H. T. (1993). An improved source model for a linear prediction speech synthesizer, Ph.D. dissertation, University of Florida.

Ishizaka, K., and Flanagan, J. L. (1972). "Synthesis of voiced sounds from a two–mass model of the vocal cords," Bell System Tech. J., 51(6), 1233–1268.

Kaplan, H. M. (1971). Anatomy and physiology of speech, McGraw–Hill, New York.

Karlsson, I. (1992). "Modelling voice variations in female speech synthesis," Speech Commun., 11, 491–495.

Kasuya, H., Ogawa, S., Mashima, K., and Ebihara, S. (1986). "Normalized noise energy as an acoustic measure to evaluate pathologic voice," J. Acoust. Soc. Am., 80, 1329–1334.

Kay, S. (1988). Modern spectrum estimation, Prentice–Hall, Englewood Cliffs, New Jersey.

Kitawaki, N. and Nagabuchi (1988). "Quality assessment of speech coding and speech synthesis system," IEEE Communication magazine, Oct., 36–44.

Klatt, D. H. (1980). "Software for a cascade / parallel formant synthesizer," J. Acoust. Soc. Am., 67(3), 971–995.

Klatt, D. H. (1986). "Detailed spectral analysis of a female voice, synthesis," J. Acoust. Soc. Am. Suppl., 1, 82, S91.

Klatt, D. H. (1987). "Review of text–to–speech conversion for English," J. Acoust. Soc. Am., 82(3), 737–793.

Klatt, D. H., and Klatt, L. C. (1990). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am., 87(2), 820–857.

Kreiman, J., Gerratt, B. R., and Berke, G. S. (1994). "The multidimensional nature of pathological vocal quality," J. Acoust. Soc. Am., 96(3), 1291–1302.

Kreiman, J., Gerratt, B. R., Precoda, K., and Berke, G. S. (1992). "Individual differences in voice quality perception," J. Speech Hear. Res., 35, 512–520.

Krishnamurthy, A. K., and Childers, D. G. (1986). "Two–channel speech analysis," IEEE Trans. Acoust., Speech, Signal Process., 34(4), 730–743.

Kuwabara, H. (1991). "Acoustic parameters of voice individuality and voice–quality control by analysis–synthesis method," Speech Commun., 10, 491–495.

Kuwabara, H., and Ohgushi, K. (1984). "Experiments of voice quality of vowels in males and females and correlation with acoustic features," Language and Speech, 27, 135–145.

Ladefoged, P. (1971). Preliminaries to linguistics phonetic, University of Chicago, Chicago, Illinois.

Lalwani, A. L. (1991). Flexible formant synthesizer: a tool for improving speech production quality, Ph.D. dissertation, University of Florida.

Lalwani, A. L., and Childers, D. G. (1991a). "A flexible formant synthesizer," Proc. IEEE Int. Conf., Acoust., Speech, Signal Process., 2, 777–780.

Lalwani, A. L., and Childers, D. G. (1991b). "Modeling vocal disorders via formant synthesis," Proc. IEEE Int. Conf., Acoust., Speech, Signal Process., 1, 505–508.

Larar, J. N., Alsaka, Y. A., and Childers, D. G. (1985). "Variability in closed phase analysis of speech," Proc. IEEE Int. Conf., Acoust., Speech, Signal Process., 1089–1092.

Laver, J. (1980). The phonetic description of voice quality, Cambridge University Press, Cambridge, New York.

Laver, J., and Hanson, R. (1981). "Describing the normal voice," in Evaluation of speech in psychiatry, edited by J. Darby, Grune and Stratton, New York, 51–78.

Lee, K. (1992). Pitch synchronous analysis/synthesis using the WRLS–VFF–VT algorithm, Ph.D. dissertation, University of Florida.

Levinson, S. E., and Schmidt, C. E. (1983). "Adaptive computation of articulatory parameters from the speech signal," J. Acoust. Soc. Am., 74(4), 1145–1154.

Linde, Y., Buzo, A., and Gray, R. M. (1980). "An algorithm for vector quantizer design," IEEE Trans. on Com., 28(1), 84–95.

Linggard, R. (1985). Electronic synthesis of speech, Cambridge University Press, Cambridge, New York.

Markel, J. D. (1972). "The SIFT algorithm for fundamental frequency estimation," IEEE Trans. on Audio and Electroacoustics, AU–20(5), 367–377.

Markel, J. D. (1973). "Application of a digital inverse filter for automatic formant and F0 analysis," IEEE Trans. on Audio and Electroacoustics, AU–21(3), 149–153.

Markel, J. D., Gray, A. H. (1974). "A linear prediction vocoder simulation based upon the autocorrelation method," IEEE Trans. Acoust., Speech, Signal Process., 22(2), 124–134.

Markel, J. D., Gray, A. H. (1976). Linear Prediction of Speech, Springer–Verlag, New York.

Makhoul, J., Roucos, S., and Gish, H. (1985). "Vector quantization in speech coding," Proc. IEEE, 73(11), 1551–1588.

Marple, S. L. (1987). Digital spectral analysis with applications, Prentice–Hall, New Jersey.

Matsumoto, T., Hiki, S., Sone, T., and Nimura, T. (1973). "Multidimensional representation of personal quality of vowels and its acoustical correlates," IEEE Trans. Acoust., Speech, and Signal Process., 21, 428–436.

McCandless, S. S. (1974). "An algorithm for automatic formant extraction using linear prediction spectra," IEEE Trans. Acoust., Speech, Signal Process., 22(2), 135–141.

Michel, J. F. and Hollien, H. (1968). "Perceptual differentiation in vocal fry and harshness," J. Speech Hear. Res., 2, 439–443.

Milenkovic, P. H. (1986) "Glottal inverse filtering by joint estimation of an AR system with a linear input model," IEEE Trans. Acoust., Speech, Signal Process., 22, 135–141.

Milenkovic, P. H. (1993). "Voice source model for continuous control of pitch period," J. Acoust. Soc. Am., 93(2), 1087–1096.

Miller, R. L. (1959). "Nature of vocal cord wave," J. Acoust. Soc. Am., 31, 667–679.

Monsen, R. B., and Engebretson, A. M. (1977). "Study of variations in the male and female glottal wave," J. Acoust. Soc. Am., 62(4), 981–993.

Murry, T. (1969). Subglottal pressure measures during vocal fry phonation, Ph.D. dissertation, University of Florida.

Murry, T. and Singh, S. (1980). "Multidimensional analysis of male and female voices," J. Acoust. Soc. Am., 68(5), 1294–1300.

Olive, J. (1971). "Automatic formant tracking in a Newton–Raphson technique," J. Acoust. Soc. Am., 50, 661–670.

Olive, J. P. (1992). "Mixed spectral representation – formant and linear predictive coding (LPC)," J. Acoust. Soc. Am., 92(4), 1837–1840.

Oppenheim, A. V., and Schafer, R. W. (1989). Discrete–time signal processing, Prentice Hall, Englewood Cliffs, New Jersey.

Pinto, N. B., Childers, D. G., and Lalwani, A. L. (1989). "Formant speech synthesis: improving production quality," IEEE Trans. Acoust., Speech, Signal Process., 37(12), 1870–1887.

Rabiner, L. R. (1968). "Digital–formant synthesizer for speech synthesis studies," J. Acoust. Soc. Am., 43, 822–828.

Rabiner, L. R., and Schafer, R. W. (1978). Digital processing of speech signals, Prentice–Hall, Englewood Cliffs, New Jersey.

Rose, R. C., and Barnwell, T. P. III (1990). "Design and performance of an analysis–by–synthesis class of predictive speech coders," IEEE Trans. Acoust. Speech Signal Process., 38(9), 1489–1503.

Rosenberg, A. E. (1971). "Effect of glottal pulse shape on the quality of natural vowels," J. Acoust. Soc. Am., 49(2), 583–590.

Rothenberg, M. R. (1973). "A new inverse–filtering technique for deriving the glottal air flow waveform during voicing," J. Acoust. Soc. Am., 53, 1632–1645.

Rothenberg, M. R., Calson, R., Granstrom, B., and Gaufin, J. (1975). "A three–parameter voice source for speech synthesis," Speech Commun., 2, 235–243.

Rothenberg, M. R. (1981). "An interactive model for the voice source," Speech Trans. Lab. –Q. Prog. Status Rep. (Royal Institute of Technology, Stockholm, Sweden), 4, 1–17.

Schroeter, J., Larar, J. N., and Sondhi, M. M. (1987). "Speech parameter estimation using a vocal tract / cord model," Proc. IEEE Int. Conf., Acoust., Speech, Signal Process., 308–311.

Schroeter, J., Larar, J. N., and Sondhi, M. M. (1988). "Multi–frame approach for parameter estimation of a physiological model of speech production," Proc. IEEE Int. Conf., Acoust., Speech, Signal Process., 88(S2.6), 83–86.

Singh, S., and Murry, T. (1978). "Multidimensional classification of normal voice qualities," J. Acoust. Soc. Am., 64(1), 81–87.

Sondhi, M. M. (1975). "Measurement of the glottal waveform," J. Acoust. Soc. Am., 57(1), 228–232.

Stephanie, S. (1986). "A computational model for the peripheral auditory system: application to speech recognition research," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 37.8.1–37.8.4.

Stevens, K. N. (1993a). "Modelling affricate consonants," Speech Commun., 13, 33–43.

Stevens, K. N. (1993b). "Models for the production and acoustics of stop consonants," Speech Commun., 13, 367–375.

Stevens, K. N., Blumstein, S. E., Glicsman, L., Burton, M., and Kurowski, K. (1992). "Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters," J. Acoust. Soc. Am., 91(5), 2979–3000.

Strum, R. D., and Kirk, D. E. (1988). First principles of discrete systems and digital signal processing, Addison–Wesley, New York.

Ting, Y. T. (1989), Adaptive estimation of time–varying signal parameters with application to speech, Ph.D. dissertation, University of Florida.

Titze, I. R. (1984). "Parameterization of the glottal area, glottal flow, and vocal fold contact area," J. Acoust. Soc. Am., 75(2), 570–580.

Trancoso, I. M., Marques, J. S., and Ribeiro, C. M. (1990). "CELP and sinusoidal coders: two solutions for speech coding at 4.8–9.6 kbps," Speech Commun., 9, 389–400.

Van den Berg, J. W. (1968). "Mechanism of the larynx and the laryngeal vibrations in fonts of phonetics," edited by J. Malmberg, North–Holland, London, 278–308.

Verhelst, W., and Nilens, P. (1986). "A modified–superposition speech synthesizer and its applications," Proc. IEEE Int. Conf., Acoust., Speech, Signal Process., 2007–2010.

Wong, D. Y., Markel, J. D., and Gray, Jr. A. H. (1979). "Least squares glottal inverse filtering from acoustic speech waveform," IEEE Trans. Acoust., Speech, Signal Process., 27(4), 350–355.

Zemlin, W. R. (1981). Speech and hearing science, Prentice–Hall, Englewood Cliffs, New Jersey.

BIOGRAPHICAL SKETCH

Yean–Jen Shue was born on March 20, 1959, in Taipei, Taiwan, Republic of China. He graduated from National Taiwan University, Taipei, Taiwan, in June 1981, with a bachelor's degree in electrical engineering. He served as a Second Lieutenant in the Marine Corps for the next two years. In 1983, he was employed by Chung Shan Institute of Science and Technology (CSIST) as a Research Assistant. During that period he worked on RF circuit design. In 1985, he got a scholarship from CSIST and started to pursue his master's degree at National Taiwan University. Two years later, he got a degree in electrical engineering; his thesis title was "Microcomputer–based automatic high vacuum monitoring system." He returned to CSIST as an Assistant Scientist and worked on the project of designing radio communication equipment and planning radio systems. In August 1991, he got scholarship from CSIST to pursue a doctoral degree and entered the graduate program of the Electrical Engineering Department at the University of Florida. He is a member of the Mind–Machine Interaction Research Center, and his research interest focuses on digital signal processing for speech synthesis and analysis. Yean–Jen is scheduled to complete his Ph.D. degree in May 1995.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Donald G. Childers, Chairman
Professor of Electrical Engineering

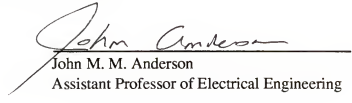
I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Frederick J. Taylor
Professor of Electrical Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Jose C. Principe
Professor of Electrical Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

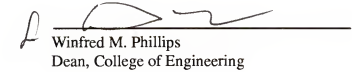


John M. M. Anderson
Assistant Professor of Electrical Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Howard B. Rothman
Professor of Communication Processes
and Disorders

This dissertation was submitted to the Graduate Faculty of the College of Engineering and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.
May, 1995



Winfred M. Phillips
Dean, College of Engineering

Karen A. Holbrook
Dean, Graduate School